Master Thesis

IU International University of Applied Sciences - Online

Study Branch: Data Science

Artificial Intelligence Driven Methods for the Analysis of Job Postings

Koen Bothmer

Matriculation Number: 92014567

Ankersweg 39

Realized in Holtum (NL)

Advisor: Prof. Dr. Tim Schlippe

Delivery date: 06-12-2021

Contents

I List of Figures

1 Introduction

As companies hiring data scientists come to realize that it becomes ever harder to find a so called "Unicorn Data Scientist" [Baškarada and Koronios, 2017], the advanced analytics field is ever specializing. With the rise of ML-Ops and Data Engineering [Mäkinen et al., 2021] roles like 'Data Engineer' and 'Machine Learning Engineer' have become reasonably well defined positions on which there seems to be consensus on responsibilities and required skills. However, there remains one position that still lacks such a consensus on the role's definition: the so-called "Data Scientist". The sheer amount of attempts to summarize the Data Scientist in Venn diagrams is illustrative for this lack of a definition [Taylor, 2016]. This rises three related questions:

- Do employers know what they are looking for in a Data Scientist?
- Do aspiring job seekers in Data Science know which skill set they should acquire?
- Do these job seekers know where to acquire this skill set?

1.1 Three Perspectives

There are three parties involved in these matters. The interaction of these parties with each other revolves around skills. Having some form of shared consensus among these parties for what the Data Scientist job entails has mutual benefits for all involved. Fig 1.1 shows the idea of a shared consensus on skills required for a job.

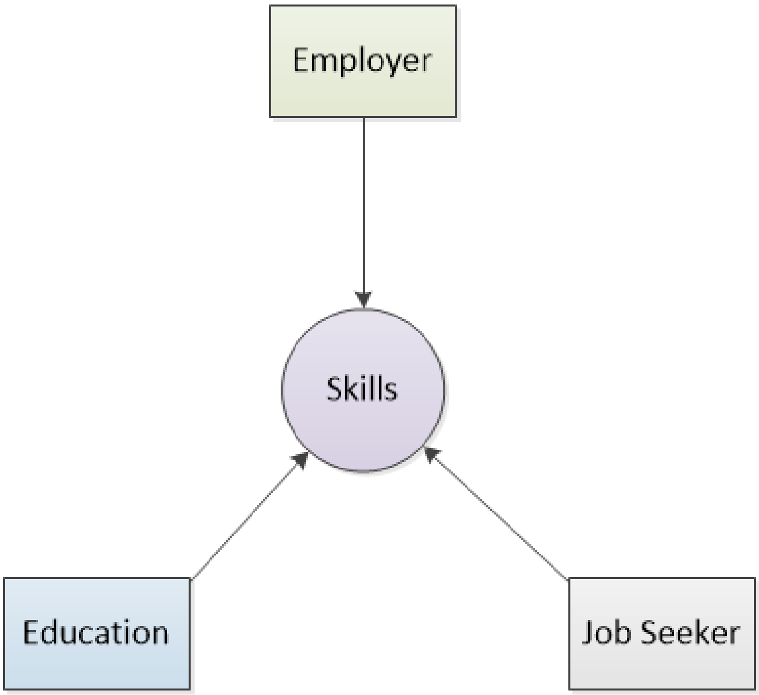


Figure 1.1: Skill Extraction from Three Sources

1.1.1 The Employer's Perspective

The skills required for a Data Scientist are driven by employers' demand; job seekers need to acquire skills so they can successfully do the work employers need them to do. Employers make this demand public in their job postings. Therefore, to define the role Data Scientist, it makes sense to try and extract a fulfilling definition from Data Scientist job postings. Such attempts have been made by [De Mauro et al., 2016], [Ho et al., 2019] and myself [Bothmer, 2021a]. These works conclude with some definition of what is and is not a Data Scientist.

Despite their defining position, Data Scientist job postings are often found to be unclear or outdated.

A shared consensus on a role's responsibilities makes it much easier for employers to find and select job seekers. It also makes it possible for employers to understand their employees' skill gaps and to select fitting education programs to fill these identified gaps.

1.1.2 The Job Seeker's Perspective

Despite the often cited high demand for Data Scientists [Davenport and Patil, 2012], Data Science proves to be a difficult field to break into. As requirements in job postings can vary wildly it is difficult for an aspiring Data Scientist to determine what learning goals to focus on. There are various online communities where these concerns are discussed daily, i.e. [Reddit, 2021]. Job seekers would benefit greatly from a clear understanding of the most important skill sets to acquire. It will help them select fitting education so that they can fit their skill set to employers' demand.

1.1.3 The Education's Perspective

79% of Data Scientist job positions require at least 3 years of experience [Miller and Hughes, 2017]. This seems to be a sign of a discrepancy between Data Science educational programs and employers' demand. A common consensus on employers' demand could help educational institutes in designing study programs that are well adjusted to the demands of the job market. A clear definition of what students will learn and how it will prepare them for the job market could also help educational institutes in advising and attracting the right students.

1.2 Overall aim

There seems to be a missing link among these three perspectives that can help them reap the suggested benefits. This work proposes a data driven approach to fill this void. The overall aim of the thesis is to gain understanding of the most commonly required skills for a Data Scientist by the analysis of job postings. This analysis provides a sense of consensus on what skills are most commonly required by employers. With this, some clarity was provided on the demand from the employer's perspective.

From the job seekers perspective it was investigated whether it is possible to compare the skills on a CV to the common requirements by employers. In the same manner, from the education point of view, learning programs were matched versus employers' demands.

The developed methodology has multiple use cases for which a prototype app was developed. This app was tested by a representative sample of respondents from each perspective. These respondents undertook a questionnaire which provided opinions on the use cases with which the work is concluded.

1.3 Structure of the Thesis

This work is best understood as one iteration of a feedback loop as it was illustrated in figure 1.2. The numbers in each block refer to the chapters of the thesis.

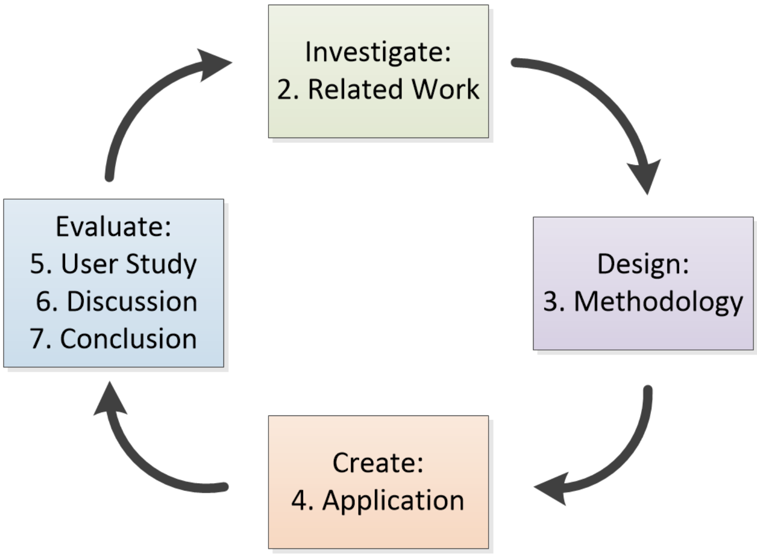


Figure 1.2: The Structure of the Thesis as a Feedback Loop

- Investigate: After this introduction a summary of related work is described in chapter 2. In the feedback loop analogy this chapter can be viewed as investigative: What is the state of AI in the field from the three aforementioned perspectives. What solutions already exist and how can this be synthesized to provide meaningful insights to each of the perspectives involved? This chapter is about generating ideas.

  The main research question this chapter seeks to answer is:

  > How is AI currently being used to improve interaction among job seekers, employers and education?

- Design: Chapter 3 explains the methodology that was developed to find consensus on what skill sets are most important to employers. Is it possible to model the demands of employers? In this chapter the main ideas from chapter 2 are taken through all the necessary steps: Data Acquisition, Data Pre-Processing, Modelling and Evaluation.

  The main research questions for this chapter is:

  > What are the most commonly required skills for a data scientist from the analysis of job postings? Can these skills be modeled so that they can be compared to skills extracted from different sources?

- Create: Based on the methodology an application was designed, developed and deployed as described in chapter 4.

  This chapter provides an answer to the question:

  > How can a skill model be applied as a service and be beneficial to job seekers, employers and educational institutes?

- Evaluate: The developed app was subjected to a user study as described in chapter 5. The feedback received from the user study provided several considerations and opportunities for future related work, hence the arrow pointing back to the investigate phase of the feedback loop. This is discussed in chapter 6 . All though not an end to the efforts of the author on this topic, in chapter 6.3 the thesis is concluded with these considerations.

This set of chapters answers the question: What are the strengths and weaknesses of the developed application?

1.4 Software Citations

This work is heavily reliant on a lot of great open source packages. All coding was done in Python [Van Rossum and Drake, 2009] making extensive use of the packages SentenceTransformers [Reimers and Gurevych, Pandas [pandas development team, 2020], Flask [Grinberg, 2018], Numpy [Harris et al., 2020], scikit-learn [Pedregosa et al., 2011] and Plotly [Inc., 2015].

2 Related Work

2.1 Use of AI based Technology in Human Recourse Management

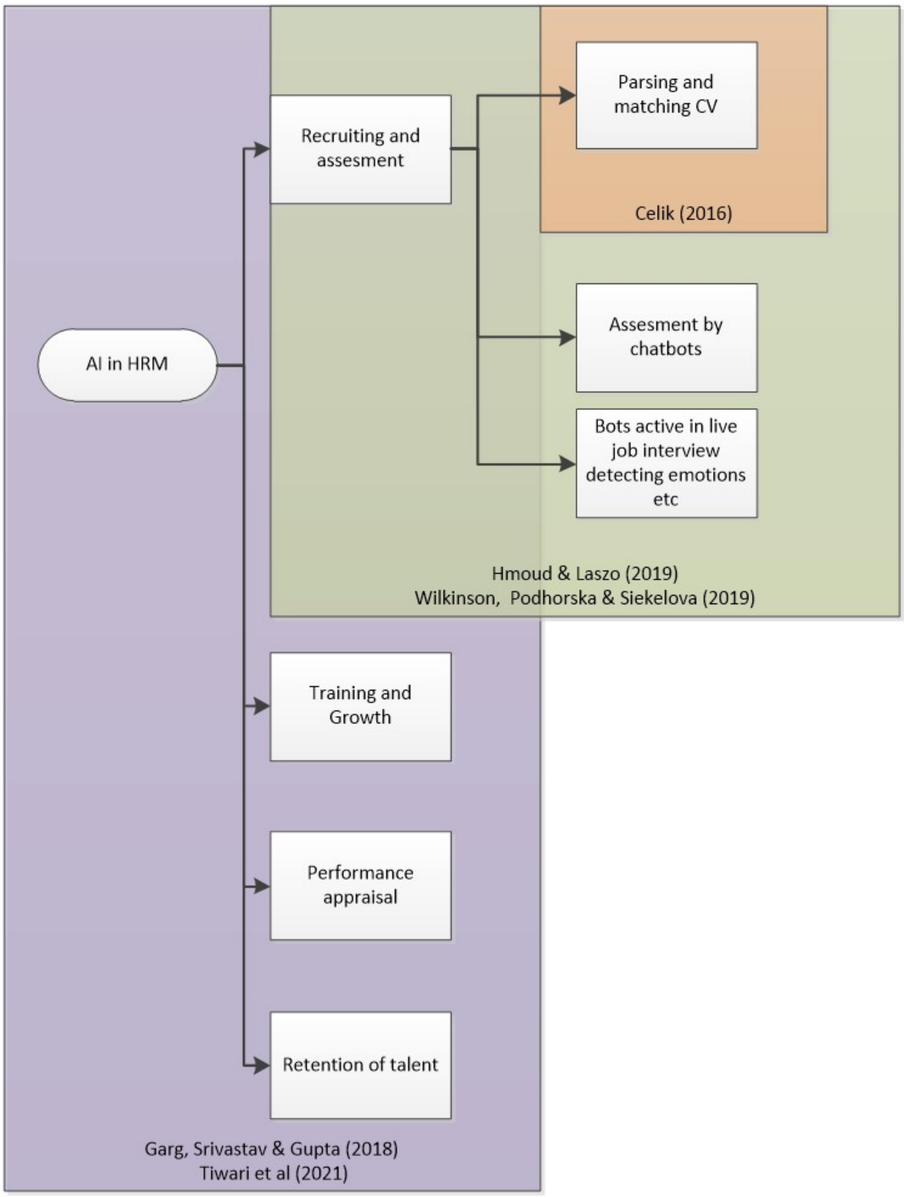


Figure 2.1: An Overview of Literature on AI in Human Resource Management

The hiring and recruitment field is not a front running field in the adoption of data and AI culture; in example, in 2019 only 14% of companies indicate they had mostly/completely adopted "Artificial Inteligence" in their practices [Wilkinson et al., 2019]. The field seems somewhat scattered which is also reflected in the describing literature as summarized in figure 2.1. [Tiwari et al., 2021] Gives a general overview of the state of AI in HRM practices. [Garg et al., 2018] Argues the use of AI in HRM has a positive environmental impact as it can cut the use of resources. [Hmoud et al., 2019] focuses on AI in recruitment processes and gives a good overview of the current state of AI in practices like interviews and assessments. [Wilkinson et al., 2019] provides insight on current perception of AI based practices in human resource management based on a large user survey. [Celik, 2016] Gives a very detailed description of skills extraction from CVs using an ontology based methodology.

There are a lot of scattered ideas and proof of concepts in this growing field but there seems to be a lack of a set of best practices regarding AI in human resource management. [Johansson and Herranen, 2019].

2.2 Three Perspectives

This lack of synthesis of AI and Data Driven techniques in human resource management offers opportunities for improvement for those that do combine the benefits of AI for multiple use cases. This thesis aims to synthesize AI benefits of analyzing skill sets from three perspectives as shown in figure 1.1. As the idea to explicitly combine skills extracted from these three sources is unique, there is no directly related work that is comparable to the efforts described in this thesis. This makes it challenging to directly answer this chapter's research question:

> How is AI currently being used to improve interaction among job seekers, employers and education?

The conducted literature research did not provide a direct answer to this question. Therefore, an overview of the current state of research for each perspective is provided in the following paragraphs.

2.2.1 The Employer's Perspective

Employers publish their needs as function requirements in their job postings. Therefore, job postings have long been an obvious source of insight in employer's demands. This data source has been analyzed time and time again for this purpose [Bennett, 2002]. No exception to this is the case of Job postings for Data Scientist jobs. There are various examples of cases in which Data Scientist job postings have been used to create an employers' demand driven definition of what the role entails, e.g. [De Mauro et al., 2016], [Ho et al., 2019].

For employers, the ability to automate the task of ranking a large number of CVs to a job description is a valuable asset. There are proposals of methodologies to achieve this, [Faliagka et al., 2014] ranks candidates for a job based on semantic matching of CV and job description. More recent natural language processing techniques offer opportunities to improve on these methods: [Fernández-Reyes and Shinde, 2019] proposed using Word2Vec to match CV's to jobs based on average word embedding. [Celik, 2016] proposed an ontology based method to extracts skills from a CV.

Recently, the natural language processing field has been shaken by the publication of BERT [Devlin et al., 2018]. Surprisingly little work using BERT for the modelling of skills has been proposed. [Nigam et al., 2020] shows a methodology that uses a BERT-based model for skill classification.

All though word and sentence embedding models are inherently unsupervised learning techniques, these works all leverage these embedding models for feature engineering to support methodologies from the supervised learning paradigm.

These kind of methods require labeled datasets for model training and validation. Because the required skills in highly technical fields like data science are ever changing, models trained on a labeled dataset have a limited shelf life. After this they require fresh labeled datasets and attention of a data scientist to be retrained. With these considerations, from an MLops perspective [Mäkinen et al., 2021], unsupervised learning approaches offer a lot of benefits as they can be retrained on the input data itself, without requiring labeling. The retraining of such models can be automated making it much easier to accommodate for the expected model drift caused by the variable nature of the required skills.

This thesis expands on the related work from the employer's perspective by proposing an unsupervised approach to skill comparison. The developed methodologies are also a lot more automatable and scalable than the methodologies employed in related work.

2.2.2 The Job Seeker's Perspective

While employers make heavy use of systems that recommend the best CV's for their jobs, from a job seeker's perspective it is interesting to recommend the best jobs to a certain CV. These so-called job recommendation systems have attracted quite a lot of research attention in the early years of the past decade [Siting et al., 2012], [Al-Otaibi and Ykhlef, 2012], [Hong et al., 2013].

Recommendation systems need a strong user profile of the job seeker in order to match people to jobs. Therefore it makes sense to deploy these systems to social media networks like LinkedIn and Facebook [Diaby et al., 2013]. As is to be expected, the job recommendation discussion seems to have moved from scientific circles to large tech companies that have the user data to build and use proper job recommendation systems.

Recent advances in natural language processing present opportunities to improve these recommendation systems. A telling sign of who is pushing these advancements in job recommendation systems is that LinkedIn's researchers are actually publishing their results on these matters [Guo et al., 2019].

All though they are not this open on all of their work, it must be noted that companies like for example LinkedIn, are showing strong recommendation based products on their apps and websites. For example, LinkedIn's website is showing features that make strong job recommendations and also learning content suggestions.

2.2.3 The Education Perspective

The idea of bilateral recommendation from the employer's and job seeker's perspective is certainly not novel [Malinowski et al., 2006]. This work expands on this idea by adding a third perspective, namely the education perspective. [Heijke et al., 2003] shows the importance of a good match between vocational competencies learned in education and an occupation to which these competencies fit. [Yorke and Knight, 2006] gives guidance in how educational programs can be designed with employability in mind. Educational institutes identified skill sets in close collaboration with relevant employers [Cox and King, 2006].

Educational institutes seem hesitant to adopt AI-based technologies. Seemingly because of a lack of faith in the educational value of such technologies [Luckin and Cukurova, 2019].

There are promising advancements like [Saito and Watanobe, 2020] who recommend a learning path based on the outcome of educational tests. [Rivera et al., 2018] performed a systematic mapping study on the use of recommendation systems in education. One of their conclusions:

> Moreover, this work has also been useful to detect some research gaps and key areas where further investigation should be performed, like the introduction of data mining and artificial intelligence in recommender system algorithms to improve personalization of academic choices. [Rivera et al., 2018]

The methodologies proposed by this thesis aim to fill these gaps.

2.3 Synthesis

Considering the state of research on the application of AI in the analysis of skills from each of these three perspectives, an image emerges of three fields that struggle independently with their own challenges. Even though the fields are undeniably related.

This work argues that all three parties benefit from consensus on which skills are important for a Data Scientist. It makes sense to look for such a consensus in job postings, because skill development is mainly an employer driven endeavor; the job market determines which skills job seekers should acquire and which skills educational institutes should teach.

This work provides a methodology to extract a skill set consensus for one job title: "Data Scientist". The methodology was designed to be as scalable as possible, making it relatively easy to expand the methodology to other job fields.

2.4 Reproducibility

An opinion of the author and many other researchers is that code and acquired data should be shared wherever possible [Peng, 2015]. As researchers are not software engineers, an often cited reason for them to not share their code is the complexity of software dependencies [LeVeque, 2013]. Because researchers do not always understand the particularities of virtual environments and environment management they often struggle to get code from one researcher to work on another researcher's computer.

A relatively simple solution to this problem is to virtualize the whole coding environment in a Docker container [Boettiger, 2015]. When sharing the code, the container can be shipped with it. Reproducibility is not a topic of this thesis but in the spirit of open research, the whole work is made reproducibly available from the provided Docker container that can be spun up following the instructions on the Github repository supporting this thesis. [Bothmer, 2021c].

3 Methodology

This chapter describes the techniques and methodology used to acquire, preprocess and model the required data and the deployment of this model in our application 'Skill Scanner'. This chapter describes the first three sections as shown in figure 3.1 in the corresponding sections. Section 4: "App" is described in chapter 4.

This methodology is heavily based on Python code. The aim of this chapter is to explain the underlying assumptions and choices made in the design of the software. For an in-depth explanation of the code itself it is best to refer to the Github repository supporting this thesis [Bothmer, 2021c]. The code was documented and commented with great care in order to enhance the reproducibility of this work.

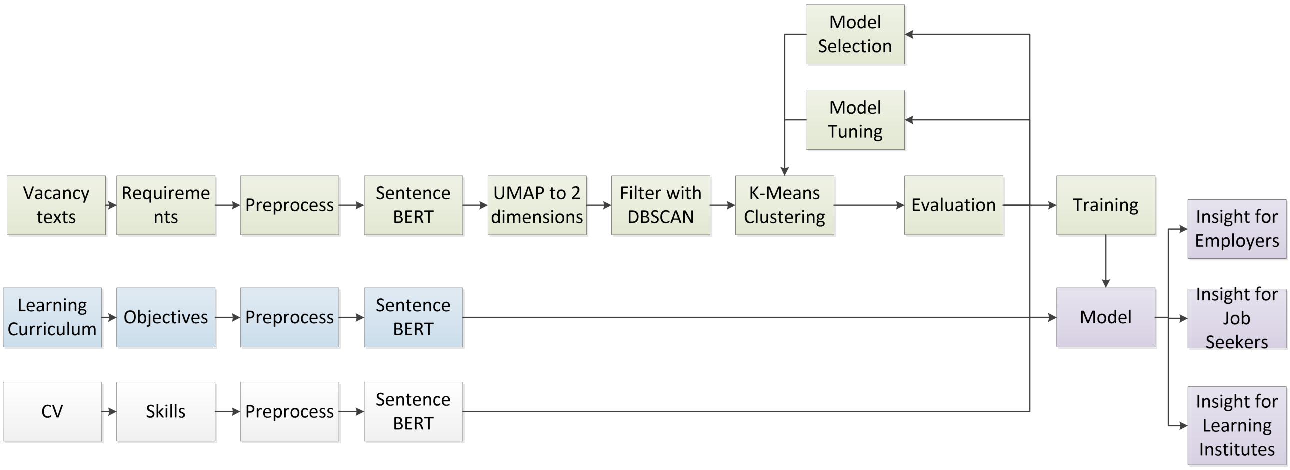


Figure 3.1: Methodology Overview

3.1 Data Acquisition

3.1.1 Acquiring Job Postings

A dataset of 1637 Western European job postings was acquired from Indeed [Indeed, 2021]. For this purpose a web scraping tool was developed using Python and the BeatifulSoup package. BeautifulSoup enables the extraction of a website's html source code [Richardson, 2007]. The web scraping tool takes in a job title and returns a dataframe of job postings that can be used for further analysis. In the case of this thesis all job postings included the bigram "Data Scientist" in their title.

To get a sample that is representative for Western Europe job postings from these 14 different countries were included: Netherlands, Germany, Belgium, Luxembourg, United Kingdom, Switzerland, Sweden, Norway, Spain, Italy, France, Portugal, Finland and Austria.

The dataset was enriched with a large dataset of USA "Data Scientist" job postings found at Kaggle [JobsPikr, 2019]. This dataset contains 10.000 job postings of which 1599 job postings could be processed into the original dataset.

The combined full dataset contains 2633 "Data Scientist" job postings extracted from the USA and Western Europe.

3.2 Data Pre-Processing

While skill extraction is a sub-field of it's own [Celik, 2016], for this work a rather simple technique was employed: In previous work of the author [Bothmer, 2021a] it was shown that elements of unordered html lists (<ul>) in job postings main texts are likely to be job requirements. The simplicity of this technique has an obvious downside: Some job postings do not have their job requirements in unordered lists, the information in those job postings is lost.

The upside of this simplistic technique is perhaps less obvious: it allows for automation of the extraction of skills from job postings. A lot of related work makes use of expert judgement at some point down the line [Debortoli et al., 2014][De Mauro et al., 2016][Keuren et al., 2020]. In this work this is not needed allowing for our methodology to be easier to generalize to other job domains. A more sophisticated skills extraction technique is beyond the scope of this work but could certainly be an improvement to the employed methodology.

BeautifulSoup [Richardson, 2007] was used to extract said list elements from the 2633 job postings in order to build a dataset of 21.509 job requirements.

3.2.1 Vectorization of Requirements

In order to computationally compare the extracted job requirements a numerical representation of said requirements in a feature space is needed. Word and sentence embedding is a fundamental sub-field of natural language processing. The purpose of these embedding techniques is to generate feature vectors that represent words or sentences as real-valued vectors in a feature space [Mikolov et al., 2013]. Words or sentences that are close in vector space are expected to be similar in meaning. For the task of comparing job requirements extracted from vacancy texts various of these techniques were tried:

3.2.2 Bag of Words

The bag of words techniques is a feature extraction technique that represents sentences and documents as a histogram-like vector of presence of words [Zhang et al., 2010]. The algorithm makes a list of all words present in a set of documents or sentences and returns a vector with the count of each word. In example:

Sentence 1: "At least 5 years of experience in programming with Python."

Sentence 2: "A minimum of 2 years of experience in a research role."

The list of all words in these sentences: [at, least, 5, years, of, experience, in, programming, with, Python, A, minimum, 2, research, role].

The vector representation of these sentences is:

Sentence 1: [1,1,1,1,1,1,1,1,1,1,0,0,0,0,0]

Sentence 2: [0,0,0,1,1,1,1,0,0,0,2,1,1,1,1]

Downsides of Bag of Words: Although a simple to understand example of a vector representation of a sentence, this technique has a lot of downsides:

- As the example shows, these two example sentences have a large similarity based on this approach. In the context of finding differences and similarities in job requirements, the vector representations of these sentences should be different for any data mining technique to work. The bag of words technique is too simplistic for these purposes as it does not pick up on the semantic meaning of different sentences beyond word count.

- Each word present in the corpus of sentences is used as a feature. In the example of 2 short sentences this resulted in a reasonable vector of 15 features. For a corpus of 21.509 job requirements the feature space will explode to a very high dimensionality. The challenges of high dimensionality are well understood and commonly referred to as "The Curse of Dimensionality" [Indyk and Motwani, 1998]. A thorough discussion on dimensionality is beyond the scope of this work but the sparsity of data points in the created feature space makes bag of words an infeasible technique to compare the extracted job requirements; all sentences will be far apart and similarities will be dificult to detect.

- The order of words in a sentence can have a large impact on the semantic meaning of a sentence. For a bag of words approach this makes no difference. Consider the next (admittedly somewhat forcefully crafted) example which illustrates this effect:

  Sentence 1: "Must have: 2 years of experience with Python, nice to have: SQL experience."

  Sentence 2: "Must have: 2 years of exprerience with SQL, nice to have: Python experience."

  When a Bag of Word based algorithm would compute a vector for these sentences, both vectors would be equal. Despite the important semantic difference imposed by the order of the words.

It should be clear from the above explanation that bag of words is not the best approach for understanding differences and similarities in job requirements.

3.2.3 Word2Vec and Glove

In 2013, the development and publication of Word2Vec was a breakthrough in the natural language processing capabilities of computers. Word2Vec algorithms use a shallow neural network architecture to create vector representations of words [Mikolov et al., 2013]. There are different model architectures:

Continuous Bag of Words: The algorithm creates a sliding window and learns to predict each word based on it's surrounding words. The model represents each word as a feature vector. The learned vectors are the embeddings the model returns.

Skip-Gram: Skipgram works the other way around, the model learns to predict the context (surrounding words) based on one word as input.

The main take-away is that the output word embedding vectors of the model are similar for words that occur in similar contexts.

Glove is a similar approach but expands on the Word2Vec algorithms by taking into account the frequency of co-occurance of words [Pennington et al., 2014]. The embeddings relate to the probability of words to appear together.

Although the strategies of these algorithms vary somewhat, they are known to achieve results of comparable quality [Naili et al., 2017].

3.2.4 An Example of Glove Embedding

Word embedding models like Glove perform best when trained on a very large corpus of training texts. That is why there are a lot of pre-trained models available.The SentenceTransformers library [Reimers and Gurevych, 2020], which was used extensively later in this thesis, offers a convenient API which allows access and use of pre-trained embedding models. This API was used to call the pre-trained model Glove 6B, which was trained on the full corpus of Wikipedia texts. This model was used for the examples in this section.

The Glove model was used to embed the word "Teamwork". Here is an excerpt of the vector representation it returns:

[-0.10511 , 0.52572 , -0.16606 , -0.52095 , 0.29966 , ... , -0.37525 , -0.20372 , -0.14162 , -1.0346]

This vector representation is not very useful on it's own. Inspired by [Alammar, 2021] colored heatmaps were created to visualize the word embedding vectors. This technique gives some intuition as to how these vector representations work.

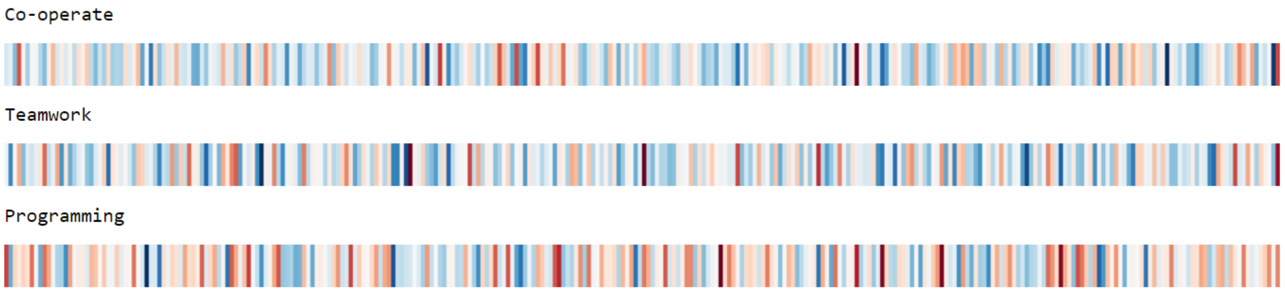


Figure 3.2: A Few Examples of Word Embedding Vectors Visualized by Color

As figure 3.2 shows, the terms "Co-operate" and "Teamwork", which are semantically similar, show a much more comparable color scheme than the term "Programming" which is unrelated to the other words. These color similarities as shown in fig 3.2 are useful because they represent numeric similarity of a large proportion of the vectors features. This means that, when mapped to a feature space, vectors that are closely related in semantic meaning, will be close to each other. Vectors of different meaning will be far apart.

3.2.5 Average Word Embeddings

The overall aim of this thesis is to find differences and similarities in job requirements, which are typically sentences. To make these sentences comparable a vector representation of the sentence is needed, word embeddings are usually averaged to achieve this.

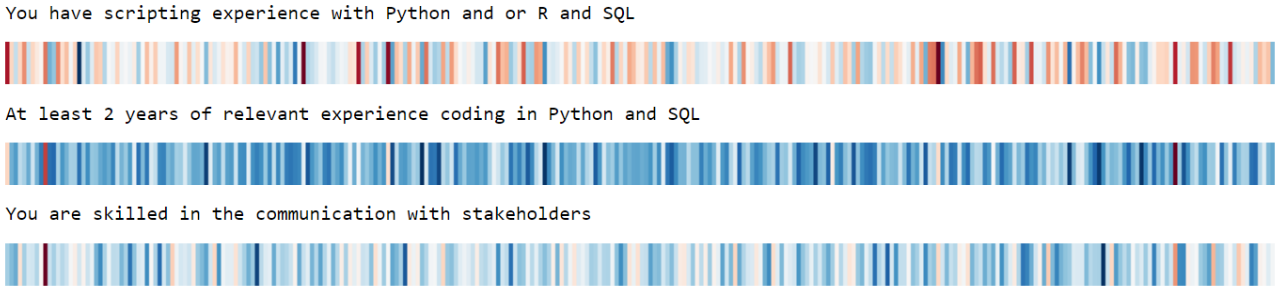


Figure 3.3: A Few Examples of Average Word Embedding Vectors Visualized by Color.

As figure 3.3 shows, the average word embedding technique gives a weak result for the use case of this work. The first two sentences, which are semantically similar in this context show very different results. The averaging of the word embedding vectors seems to distort the result too much to retain sufficient meaning to make job requirements comparable. Presumably, this is caused by the fact that an average shifts towards a part of the sentence when it is described with more words. In any case, the average word embedding model seems not useful for the purpose of this thesis.

3.2.6 Doc2Vec

The averaging of word embedding vectors is a very rudimentary way of representing collections of words as vectors. More sophisticated approaches, like Doc2Vec have been proposed. Doc2Vec use an algorithm similar to the continuous bag of words (CBOW) algorithm of Word2Vec. Just like CBOW, Doc2Vec learns feature vector representations by predicting words based on their context of surrounding words. Where Doc2Vec differs is that next to the surrounding words a document feature vector is also added to the context [Le and Mikolov, 2014].

For the task of understanding differences and similarities in job requirements some experimentation with Doc2Vec has taken place. As pre-trained Doc2Vec models that are fit for the task were not available some experimentation with training a Doc2Vec model on the dataset of job requirements was undertaken. The results were unsatisfying. this is explainable because Doc2Vec and Word2Vec models require large corpusses for effective training (milions of words or even milions of documents as opposed to a meager 21.509 job requirements).

3.2.7 BERT

The most recent major breakthrough in natural language processing has come in 2018 with the publication of Google's work on BERT [Devlin et al., 2018]. Unlike Word2Vec algorithms, which move through an input text in one direction (often left-to-right), BERT learns in the bidirectional fashion of transformer models. It allows for these models to learn a deeper sense of context of natural language.

BERT-models do not compute individual sentence embeddings. This makes these models inapplicable for tasks like similarity comparison and clustering [Reimers and Gurevych, 2020].

3.2.8 SentenceTransformers

Sentence-BERT proposes a modification of the BERT network architecture using siamese and triplet networks and has the ability to compute sentence embeddings [Reimers and Gurevych, 2020]. These embeddings enable the model to be used in similarity comparison of sentences, like the use case of this thesis.

The developers of Sentence-Bert conveniently offer their methodology through an API. This API also offers a large range of pre-trained models. Because the dataset of 21.509 job requirements is too small to train our own Sentence-BERT model it makes sense to make use of these pre-trained models.

The best results were achieved using the model all-distilroberta-v1. The evaluation method used to select this model is described in section 3.3.1. The all-distilroberta-v1 is a model trained using the RoBERTa training approach which stands for Robustly Optimized BERT Pretraining Approach [Liu et al., 2019].

In the example in figure 3.4 the same sentences that were unsuccesfully embedded using average word embeddings in figure 3.3 were now embedded using SentenceTransformers' all-distilroberta-v1 model. Because this model encodes the sentences to a vector representation of 768 dimensions, it is not feasible to plot the whole vector as a heatmap. In figure 3.4 the first 100 features of the vectors are shown.

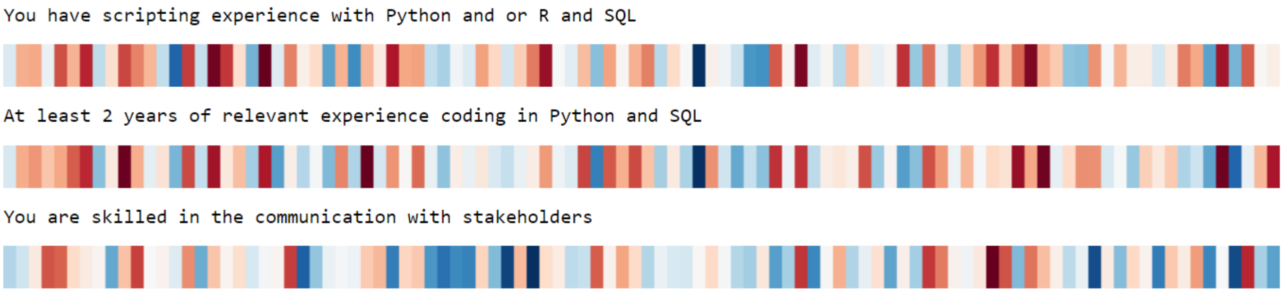


Figure 3.4: A Few Examples of Sentence Embeddings Visualized by Color

As figure 3.4 shows, the first two requirements, which are crafted to be semantically similar, show a lot of similarity in their vector representation. The third, unrelated requirement also shows a very dissimilar vector representation.

The representations shown in figure 3.4 are merely for explanatory purposes. Off course, a more rigid methodology was used for model evaluation and selection purposes. These are described in section 3.3.

3.2.9 Dimensionality Reduction

A common problem of machine learning practicioners is the impossibility of visualizing a high dimensional feature space [Indyk and Motwani, 1998]. With data visualization as the main purpose, a dimensionality reduction to a linear subspace like PCA usually proves to be sufficient for most use cases [Van Der Maaten et al., 2009]. However, in our use case of analyzing job requirements, a dimensionality reduction technique that leaves the right information intact serves a second purpose:

The dataset of job requirements was extracted from job postings by splitting unordered lists in the main text of the vacancy texts. These 'bullet points' proved likely to be job requirements in previous work of the author [Bothmer, 2021a], but they can, off course, be anything else. Given this noisy nature of the data source, there are a lot of outliers present in the dataset of job requirements.

A natural candidate to find and eliminate this noise are density based clustering algorithms. However, these algorithms are known to perform weakly in high dimensional feature space. This is caused by "The Curse of Dimensionality" [Indyk and Motwani, 1998], because distances in high dimensional feature space tend to converge, all points will be far apart. This makes it impossible to identify dense areas.

Dimensionality reduction could enable a density based approach but it relies heavily on the quality of the dimensionality reduction. Consider figure 3.5, where we show an example of applying PCA on sentence embeddings of the job requirement dataset.

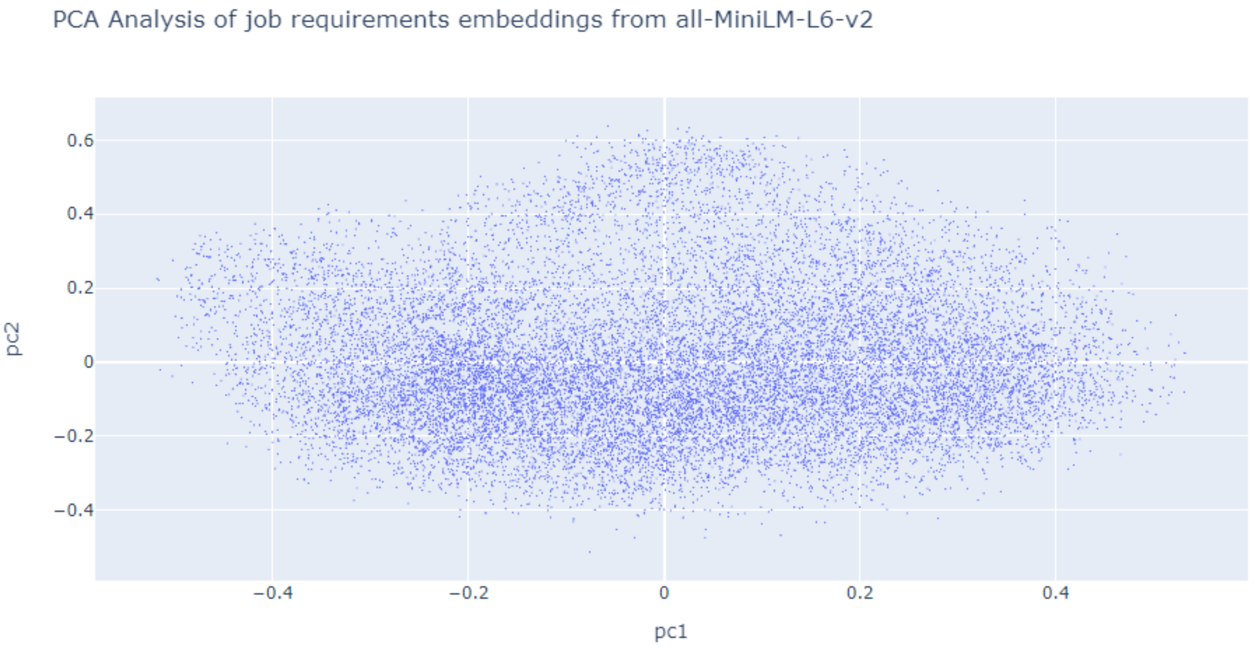


Figure 3.5: Example of PCA Dimensionality Reduction of Sentence Embedding Vectors to Two Dimensions

It should be obvious from figure 3.5 that the loss of information after applying PCA is too large to make density based clustering a feasible option for these sentence embedding vectors. A step of experimentation with non-linear dimensionality reduction techniques took place.

Some literature research appointed UMAP as a natural candidate for the use case of analyzing job requriements: The choice was inspired by the work of [Allaoui et al., 2020] which shows a remarkable improvement of performance of density based clustering algorithms by adding UMAP as a pre-processing step. More specific to the purposes of this thesis is the case of Top2Vec [Angelov, 2020], where embeddings of similar nature to our use case are preprocessed using UMAP which proved to be an enabler for the use of HDBSCAN.

UMAP [McInnes et al., 2020] is a uniform manifold learning technique proposed in 2020, it offers a novel non-linear dimensionality reduction technique based on Riemannian geometry and algebraic topology. Consider figure 3.6, where the same sentence embeddings as those shown in figure 3.5 where reduced to two dimensions, this time using the UMAP algorithm.

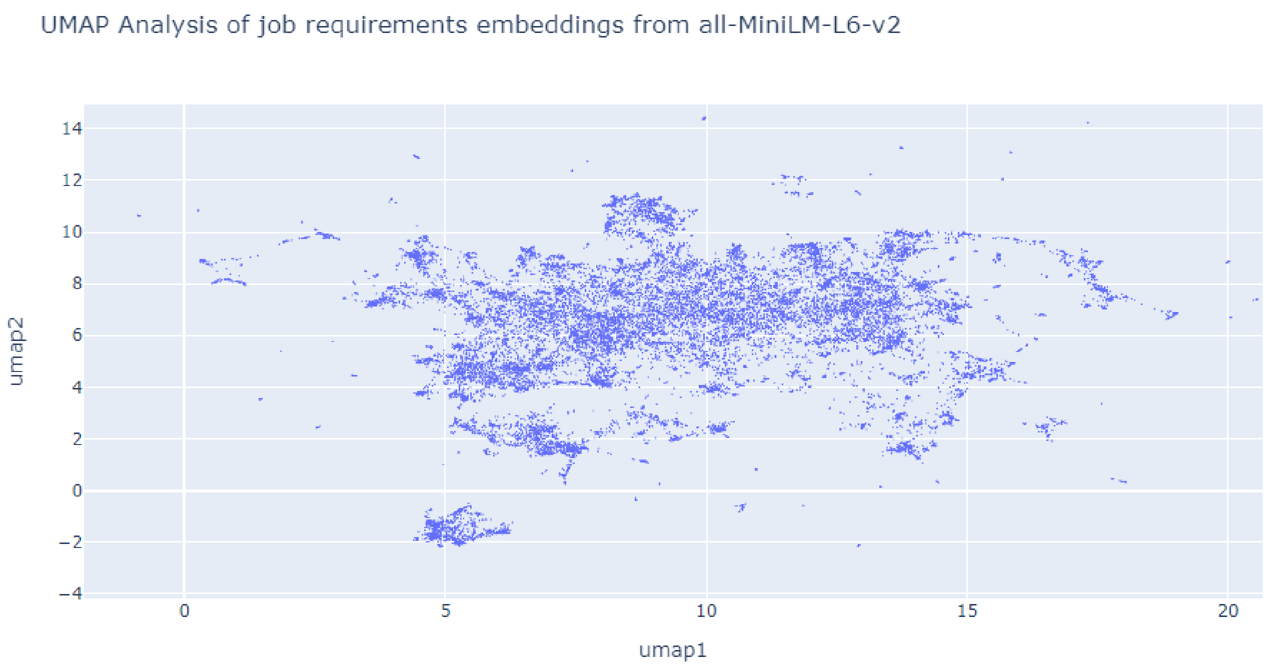


Figure 3.6: Example of UMAP Dimensionality Reduction of Sentence Embedding Vectors to Two Dimensions

The visual evidence provided by figure 3.6 and figure 3.5 should be sufficient to state that the UMAP technique is much more able than PCA to uncover patterns in the underlying data source. The data points in the 2-dimensional projection clearly show differences in densities. There are a lot of dots and flocks far out in feature space which could well be outliers that need to be removed. This dimensionality reduction technique shows potential as an enabler for density based clustering.

3.2.10 Data Cleaning with DBSCAN

UMAP was able to reduce dimensionality of sentence embedding vectors and still retain sufficient information to reveal which job requirements in the source data are very different from all other requirements. Using the right combination of sentence embedding model, UMAP and DBSCAN makes it possible to clean the dataset from outliers. Density based algorithms are known for their capabilities to remove outliers [Ester et al., 1996], in this work, DBSCAN was used for this purpose.

A range of models from the SentenceTransformers library [Reimers and Gurevych, 2020] were tested in combination with UMAP and DBSCAN. In figure 3.7 the data cleaning step is further explained.

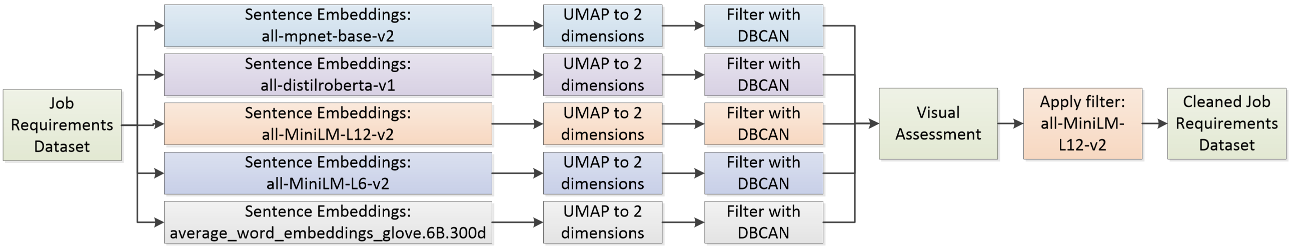
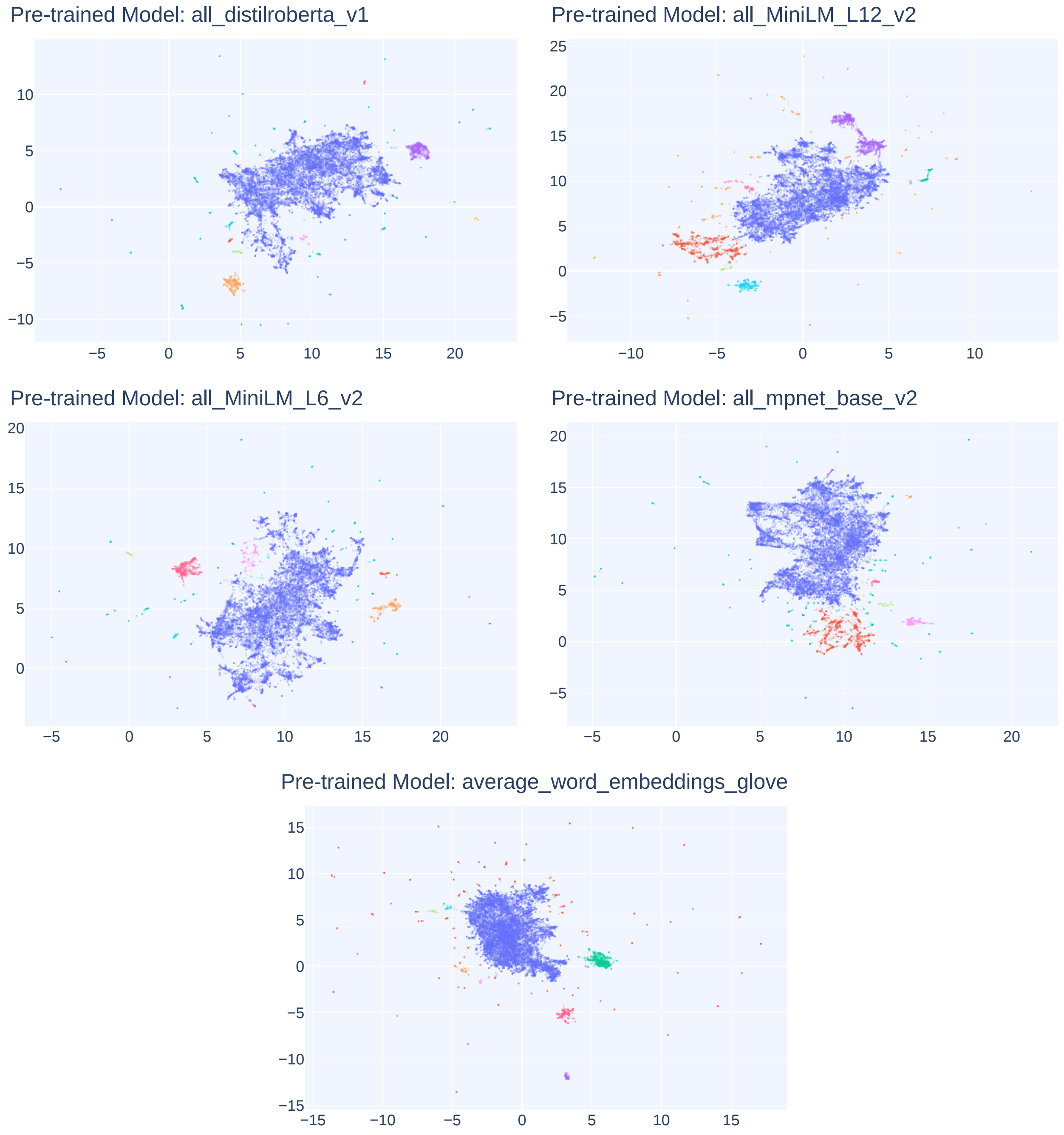

Job Requirements Dataset
Sentence Embeddings: all-mpnet-base-v2
Sentence Embeddings: all-distilroberta-v1
Sentence Embeddings: all-MiniLM-L12-v2
Sentence Embeddings: all-MiniLM-L6-v2
Sentence Embeddings: average_word_embeddings_glove.6B.300d
UMAP to 2 dimensions
UMAP to 2 dimensions
UMAP to 2 dimensions
UMAP to 2 dimensions
UMAP to 2 dimensions
Filter with DBCAN
Filter with DBCAN
Filter with DBCAN
Filter with DBCAN
Filter with DBCAN
Visual Assessment
Apply filter: all-MiniLM-L12-v2
Cleaned Job Requirements Dataset

Figure 3.7: UMAP-DBSCAN Data Cleaning Step Explained

Each of the pre-trained models shown in figure 3.7 were used to infer sentence embedding vectors for each job requirement in the dataset of 21.509 job requirements. The dimensionality of each of these results was reduced to two dimensions and DBSCAN was used to cluster the requirement UMAP embedding results. The results for each pre-trained sentence embedding model are shown in figure 3.8.


Pre-trained Model: all_distilroberta_v1

Pre-trained Model: all_MiniLM_L12_v2

Pre-trained Model: all_MiniLM_L6_v2

Pre-trained Model: all_mpnet_base_v2

Pre-trained Model: average_word_embeddings_glove

Figure 3.8: UMAP-DBSCAN Results for Each Pre-Trained Model

Figure 3.8 shows that the sentence-bert models show better results than the glove average word embedding model. The plots this figure are interactive in the notebook ./Data_Cleaning_DBSCAN Cleaning_with_DBSCAN.ipynb, available at the Github repository supporting this thesis. By hovering over the individual data points, the original job requirement texts are shown. This way, it is possible to get an idea of what each noise point and cluster contains. This technique was used to visually assess each of the results from figure 3.8 By means of this assessment, the model all_MiniLM_L12_v2 was selected as it shows the best results at detecting outliers in the data source.

Note: It must be understood that the model all_MiniLM_L12_v2 was only used for outlier reduction. A more extensive model selection step took place to select the pre-trained embedding model used for the later steps in the data pipeline. This is described in section 3.3.1.

Figure 3.9 shows the result of embedding the job requirements with model all_MiniLM_L12_v2. These embeddings were reduced to 2 dimensions using UMAP. This projection of the job requirements in 2-dimensional space was clustered by the DBSCAN algorithm. The results are shown in figure 3.9.

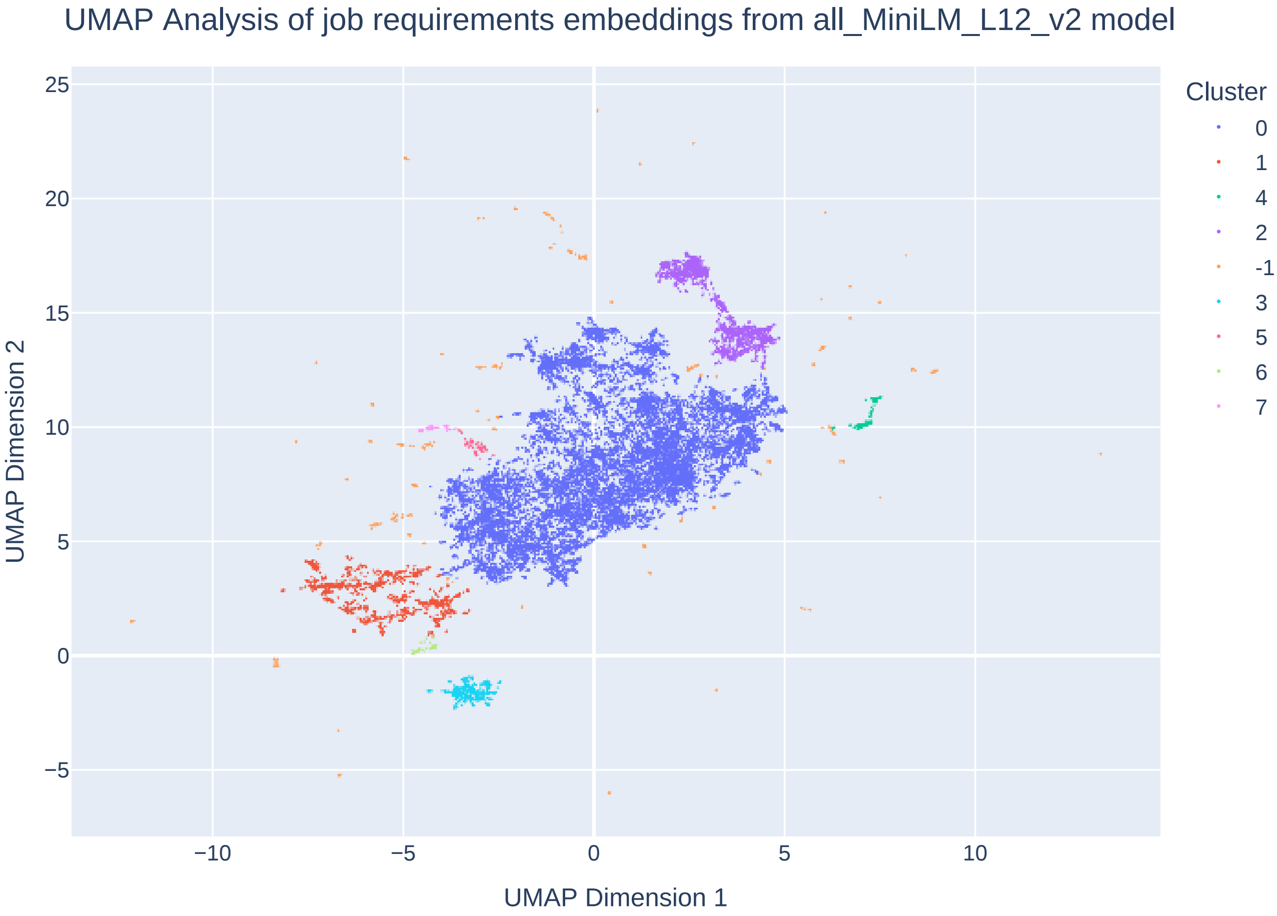


Figure 3.9: Dimensionality Reduced Embeddings Clustered by DBSCAN

3.2.11 Content Based Cleaning by DBSCAN Output

As shown in figure 3.9, the DBSCAN algorithm has clustered the data into 9 clusters. Because most clusters are located at the edges of the feature space, it was necessary to manually assess whether each of the clusters contains mostly job requirements or noise. In the notebook available at [Bothmer, 2021c] an interactive version of figure 3.9 is available. Hovering over each point shows the original requirement. There is also a function available which returns the 3 requirements closest to the center of the cluster. This function and the interactive plot were used to estimate the content of each cluster. This estimation was used to decide whether to keep or drop the cluster's content:

- Cluster 1: Dropped, Contains Job benefits and location data.
- Cluster 2: Keep, contains relevant requirements on medical subfield.
- Cluster 3: Keep, contains relevant language requirements.
- Cluster 4: Keep, contains relevant requirements on software development skills.
- Cluster 5: Dropped, contains information on the job interview process

- Cluster 6: Dropped, noise, data splitting errors caused by bad use of HTML unordered lists in some vacancy texts.
- Cluster 7: Dropped, noise on citizenship, not job specific requirements.
- Cluster 0: This is the main cluster containing most of the none-outliers. This cluster was obviously not removed.
- Cluster -1: This cluster contains all requirements DBSCAN has clustered as noise points (outside of any core-point's e-neighborhood). In order to evaluate whether these points are truly outliers we assessed the clusters using the interactive plot. When hovering over each data point the original job requirement is shown. Hovering over the points in cluster -1 shows a range of information which is almost never a job requirement. Subsequently, cluster -1 was removed. To check on this assessment, the reader is referred to the data cleaning notebook available at:

  [Bothmer, 2021c] at ./Data_Cleaning_DBSCAN/Cleaning_with_DBSCAN.ipynb.

This data cleaning step results in a dataset of extracted unordered list items that are much more likely to be job requirements. The original dataset of 21.509 list items was reduced to a subset of 18.786 job requirements.

3.3 Modelling

The main purpose of this thesis, to find differences and similarities in job requirements, is a typical unsupervised learning task. To effectively mine the dataset of 18.786 job requirements it makes sense to adopt an approach based on clustering algorithms. There are 4 main families of clustering algorithms [Verma et al., 2012]:

- Density Based Clustering: Density based clustering algorithms find clusters in feature space based on how many data points are present in close proximity to a data point. Dense areas are clustered together whereas data points in sparse areas are appointed as noise points. An example of density based clustering is DBSCAN [Ester et al., 1996], it was used earlier in this work for data cleaning purposes in section 3.2.10.

  A downside of density based clustering algorithms is that their performance is known to detoriate in high dimensional feature spaces. The typical dimensionality of embedding models that give good results is 768 dimensions. Density based clustering in a feature space of this dimensionality is infeasible. Therefore, to enable density based clustering algorithms, the dimensionality of the feature space has to be reduced. For the outlier detection step described in section 3.2.10 this worked well as outliers tend to be very different and well detectable in the 2-dimensional feature space projection of UMAP. For more sophisticated data mining of the job requirement dataset the information preserved by UMAP proved to be insufficient, returning unsatisfying results.
- Hierarchical Clustering: Hierarchical Clustering algorithms build a tree of clusters, commonly referred to as a dendrogram [Verma et al., 2012]. A hierarchical approach could make sense in the use case of this thesis; Job requirements are hierarchical by nature: in example, one could split the job Data Scientist into Hard Skills and Soft Skills, then split these into smaller and smaller subcategories.

However, hierarchical clustering approaches are very sensitive to outliers. This sensitivity is understandable: These algorithms suffer from the fact that one incorrect splitting decision distorts the result of all clusters created from these incorrectly chosen parent leafs [Rani^1 and Rohil, 2013]. Robustness to outliers is very important for this work: A data pipeline which is automated as much as possible is a priority to make this work as generally applicable as possible.

Hierarchical clustering algorithms are computationally expensive and therefore not feasible to use on large datasets. The use case of clustering 18.786 job requirements is somewhat of an edge case computationally. Adopting this approach would severely limit the possibilities to scale the products build upon this methodology.

- Gaussian Mixture Models: These clustering algorithms perform clustering of datasets by maximizing the probability of a data point to belong to a certain cluster. The computation of the probability distribution of each individual data point is computationally expensive. Just like hierarchical clustering, an approach based on Gaussian Mixture Modelling would severely limit the scalability of the developed methodology. Therefore this model family is not a feasible candidate for the use cases of this thesis.

- Centroid Based Clustering: K-means clustering, the most common of all centroid based clustering algorithms has definitely stood the test of time [MacQueen et al., 1967]. The principle of this technique is not hard to understand: The user chooses an arbitrary number of clusters K. A number of K points are randomly divided in feature space, these points are called centroids. Each data point is appointed to it's closest centroid. Now, based on the position of each data point, the centroids are recomputed as the arithmetic mean of all data points present in the cluster. This is one iteration. In the next iteration the data points are redivided over the new centroids. Once data points do not change cluster anymore, the model has converged.

  K-means clustering has been successfully used in clustering word embedding vectors [Zhang et al., 2018].

  As K-means clustering allows for adaptable and scalable solutions, the method was selected as most appropriate for the use cases of this work.

Like common practice in clustering of high dimensional data, the similarity metric that was used in all K-means models for this work is cosine similarity.

3.3.1 Model Selection

Unsupervised learning approaches are notoriously difficult to evaluate. Especially cases like the clustering of job requirements in this thesis are problematic because there is no ground truth; there are no data labels to compute commonly used evaluation metrics.

Some literature research took place in order to determine an appropriate evaluation method. In comparable work there is usually some form of ground truth for evaluation purposes:

- JobBERT [Decorte et al., 2021] uses sentence embeddings of skills to normalize job titles; they argue jobs with similar skills should be titled the same. They propose their own labeled dataset for evaluation.

- ConSultantBERT [Lavi et al., 2021], research work by Randstad, matches jobs and job seekers based on Randstad's own interaction data. When job seekers (in this case consultants) react to a job opportunity the CV is threated like a match if the consultant received reaction from the hiring company. This gives the researchers a training dataset with which classic confusion matrix based evaluation is possible.

Given the reliance on labeled data for most of the directly related work, most optimization strategies rely on the classic train-test-validate paradigm. As labeled data is hard to attain in our case a method is needed that is based on the clustering model itself. For this task, silhouette score was selected as it has been shown to be an effective optimization metric for k-means clustering approaches [Lletı et al., 2004].

Consider figure 3.10. The silhouette coefficient for one sample is computed as follows:

$$Silhouette\_Coefficient = \frac{b - \bar{a}}{max(a, b)}$$

The coefficients of all data points are averaged to compute the overall silhouette score. The best possible score is 1. A score near 0 means there are overlapping clusters present in the dataset. A negative score means data points have been assigned to the wrong cluster [Pedregosa et al., 2011].

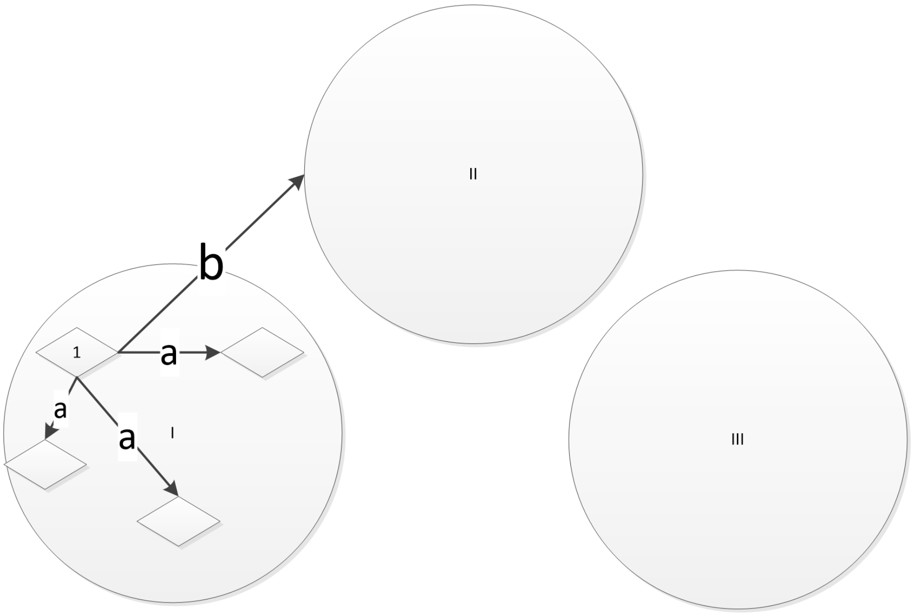


Figure 3.10: Silhouette Score Explained.

The selection of the most appropriate sentence embedding model and the tuning of the number of clusters should be performed together; different sentence embedding models will probably have a different optimal number of clusters. This is why for each sentence embedding model an array of models with different values for the number of clusters K was trained and evaluated. This is explained in figure 3.11.

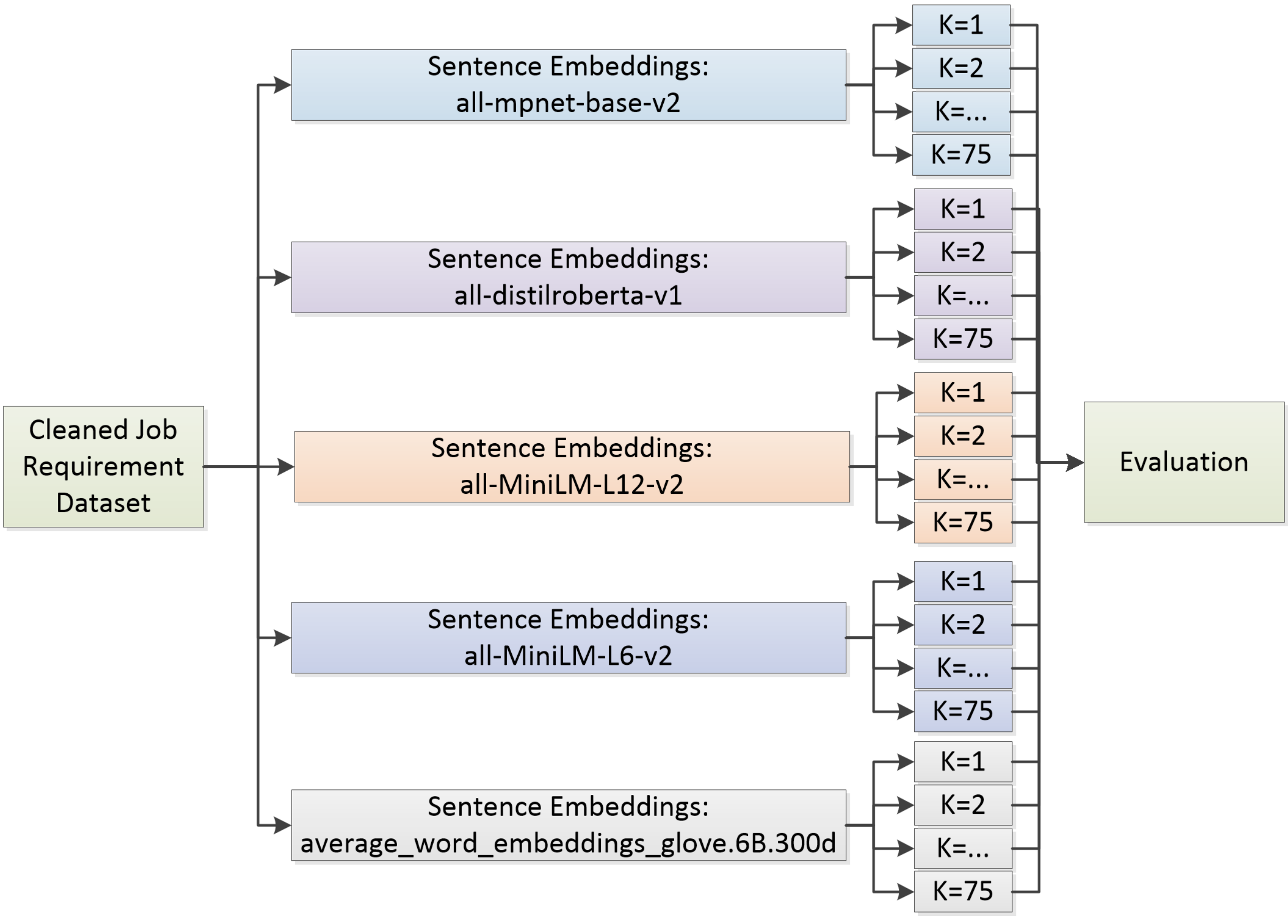

Cleaned Job Requirement Dataset
Sentence Embeddings: all-mpnet-base-v2
Sentence Embeddings: all-distilroberta-v1
Sentence Embeddings: all-MiniLM-L12-v2
Sentence Embeddings: all-MiniLM-L6-v2
Sentence Embeddings: average_word_embeddings_glove.6B.300d
K=1
K=2
K=...
K=75
Evaluation

Figure 3.11: A Grid Search for Optimal Combinations of Sentence Embedding Models and K-means Models.

The optimal model with optimal parameter k was determined using a grid search: For each embedding model shown in figure 3.11 an array of k-means clustering models was trained. These are the values for k that were evaluated: 10, 15, 20, 25, 30, 32, 34, 36, 38, 40, 42, 44, 50, 75. The average silhouette score was computed for each embedding model for each of these values for the number of clusters k.

The resulting silhouette scores are shown in figure 3.12. All though the differences are marginal (the y-axis is truncated), the model 'all_distilroberta_v1' shows to deliver the best results.

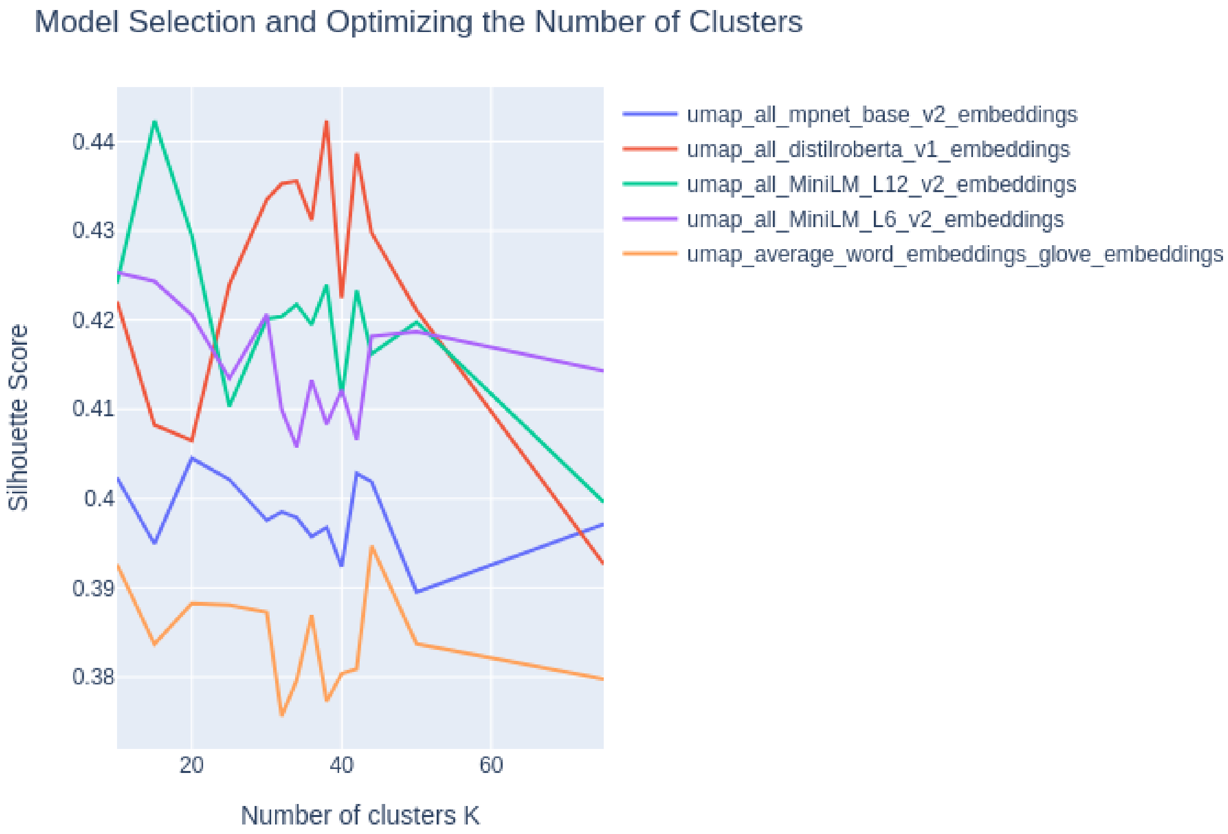


Figure 3.12: The Resulting Silhouette Score of All Evaluated Models.

As the figure 3.12 shows, for the silhouette score metric, the optimal model is "all_distilroberta_v1". The optimal number of clusters is between K=25 and K=44, for further analysis a k-means model was trained for each K between these values. The results can be seen in figure 3.13.

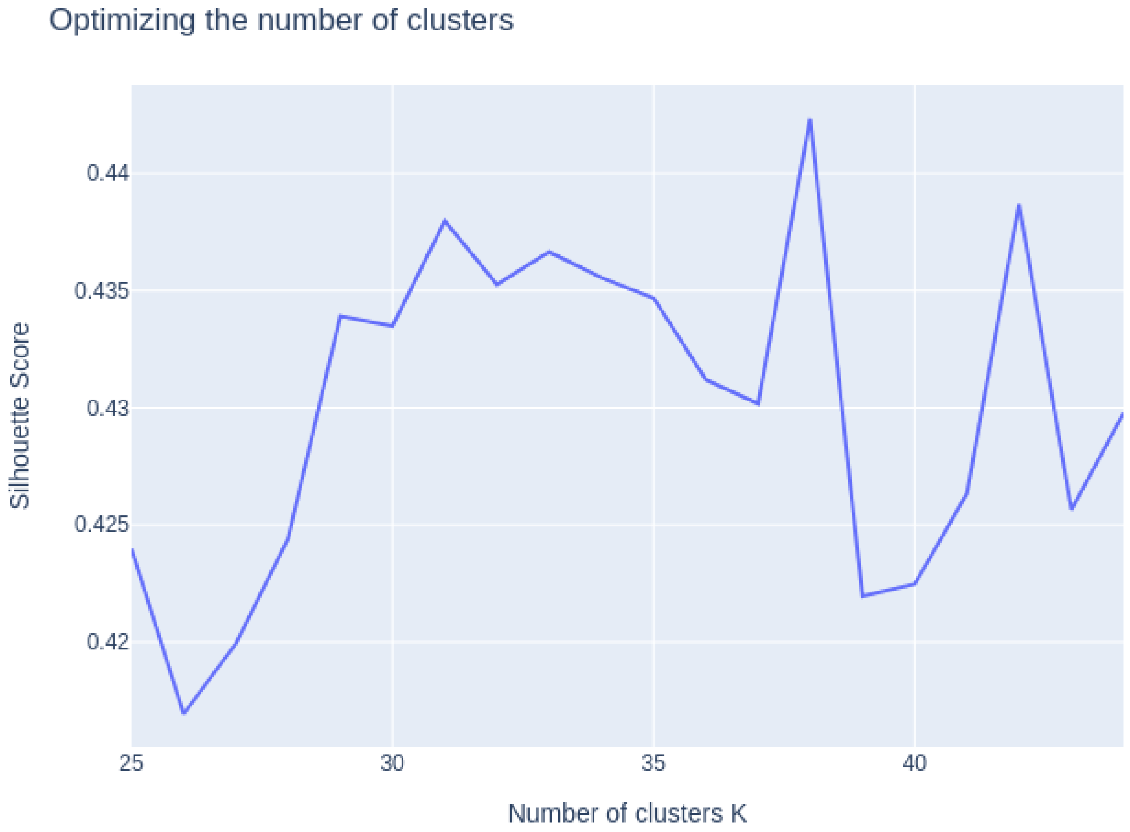


Figure 3.13: The Resulting Silhouette Score of Multiple K-means Models Trained on distilroberta Embedding Vectors of Job Requirements.

Later in the pipeline there is an optimization step, described in section 4.1. An analysis of the frequency of cluster presence in each job posting shows that a lower number for K shows is beneficial. The reasoning for this choice is further explained in section 4.1.

After these considerations, the combination of embedding model "distilroberta" with K-Means clusterer with K=31 was selected as the optimal approach for clustering the dataset of 18.786 job requirements. A UMAP projection of the result of clustering this dataset is shown in figure 3.14.

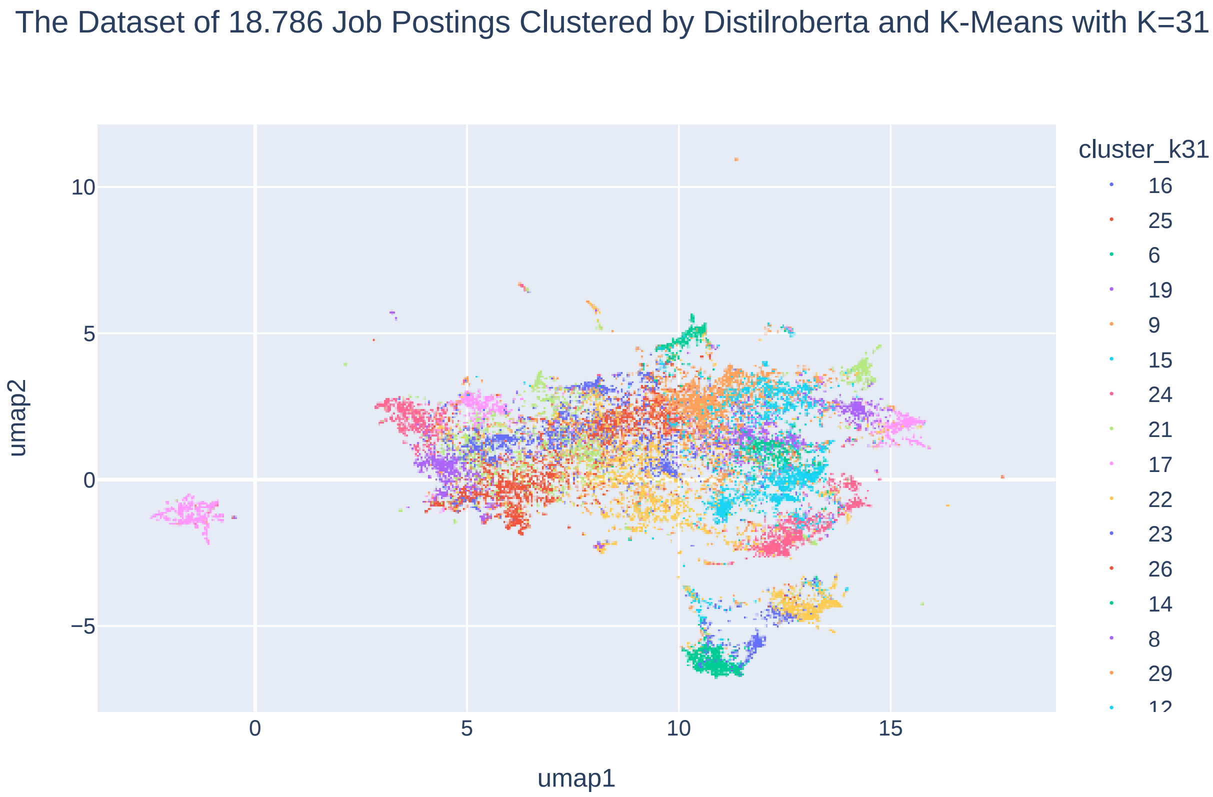


Figure 3.14: A UMAP Projection of Clustering Job Requirements.

3.3.2 Model Evaluation

As was described in section 3.3.1, in most related work there is some kind of ground truth from which confusion matrix based metrics like accuracy, sensitivity and specificity can be computed [Decorte et al., 2021] [Lavi et al., 2021]. In the case of this thesis there is no such ground truth, this makes it challenging to evaluate the model.

The difficulties of evaluating unsupervised learning approaches like clustering are generally acknowledged. [Von Luxburg et al., 2012] argues that clustering algorithms are best evaluated by taking into account the use case of the clustering approach.

Now, to adopt this idea of evaluating by use case, it is important to revisit the purpose of this thesis: Consider figure 3.15. Essentially, the methodology up until this point has described the green part of training a K-means model based on job requirements, extracted from vacancy texts. The central idea of training this model in the first place is to have a consensus based model on what employers demand of a Data Scientist. This model can subsequently be used to infer how well a learning curriculum and/or CV cover this extracted employers' demand. These routes are represented as the blue and grey parts of figure 3.15.

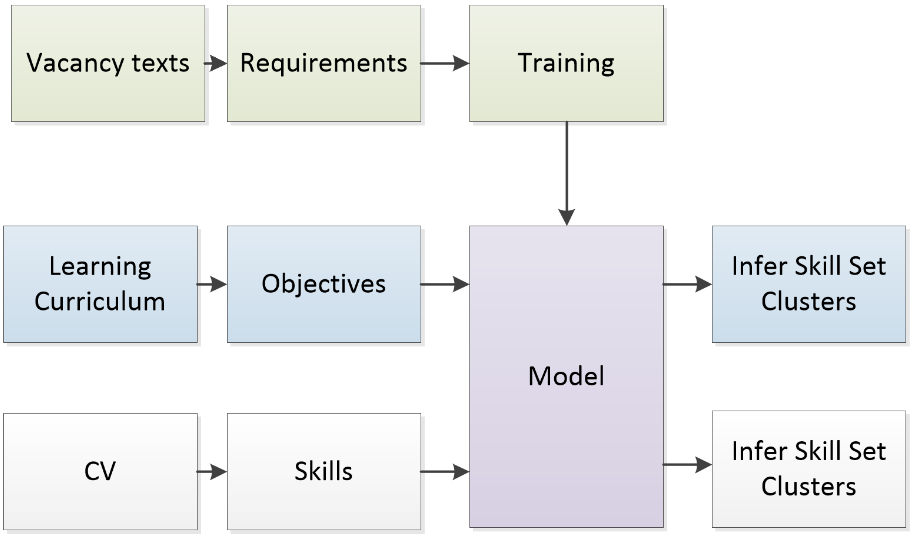


Figure 3.15: The Purpose of the Modelling Efforts.

Because the model was trained solely on vacancy text data, an important aspect to research is the ability of the model to generalize to data from learning curricula and CVs. Being able to infer the correct clusters on unseen data from the blue and grey route in figure 3.15 will provide insight in how well the model is fit for the specific use case of this thesis.

3.3.3 Evaluation by Manual Assessment of Inferred Clusters

The accuracy of the inferred clusters has to be assessed manually. In order to achieve this, a sample of Learning Curricula and CV data was used as described in figure 3.16:

- Learning Curriculum (blue route): the program "Msc Data Scientist 60ECT" of IU International University of Applied Sciences was used as a sample of a relevant learning curriculum. The module handbook proved to contain learning objectives per module. Some simple text splitting techniques showed to be sufficient to extract the learning objectives from this curriculum, code available as a notebook [Bothmer, 2021c].

  The model trained on Job Requirements was used to infer a skill set cluster for each learning objective.

  By comparing the content of the learning objective to the content of the inferred cluster the accuracy of the predicted cluster was assessed by hand as True or False.

  From this assessment the accuracy of the model to infer the correct skill set clusters of learning requirements was estimated at 77%.

- Data Scientist CVs: [Jiechieu and Tsopze, 2021] published a large dataset of 30.000 CVs from the IT domain. From these CVs a sample of 65 Data Scientist CVs was extracted. With some simple text splitting techniques a set of 5805 skills were extracted from these Data Scientist CVs, code available as a notebook [Bothmer, 2021c].

  For evaluation purposes, 100 skills were randomly sampled from this dataset of skills. The model trained on job requirements was used to infer a skill set cluster of each skill in this sample.

  The content of each sample skill was compared to the content of the inferred cluster. The accuracy of the predicted cluster was assessed by hand as True or False.

The accuracy of the model to infer skill set clusters of CV skills was estimated at 88%.

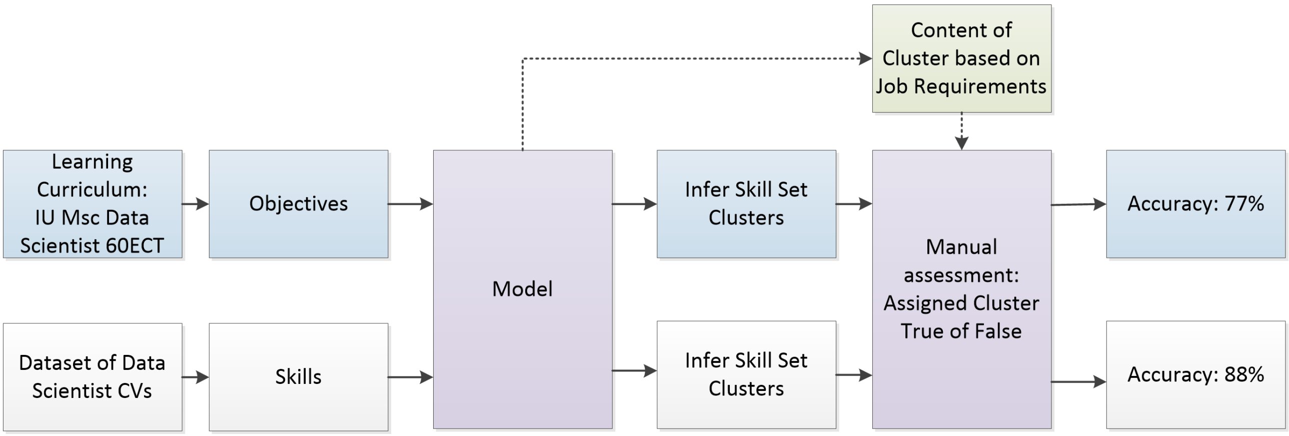

Content of Cluster based on Job Requirements
Learning Curriculum: IU Msc Data Scientist 60ECT
Objectives
Model
Infer Skill Set Clusters
Manual assessment: Assigned Cluster True of False
Accuracy: 77%
Dataset of Data Scientist CVs
Skills
Infer Skill Set Clusters
Accuracy: 88%

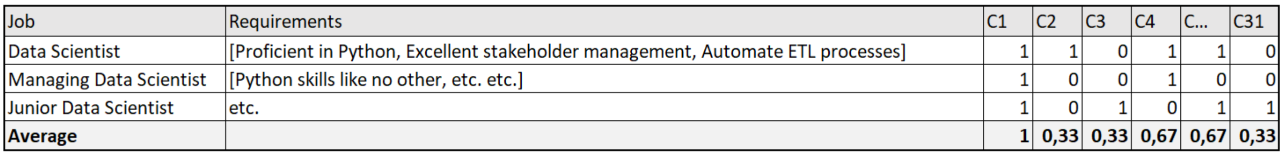
Figure 3.16: Manual Evaluation: Experiment Set-up.

4 Application

As was demonstrated in section 3.3.3, the modelling pipeline is able to cluster data from CVs and learning programs with reasonable accuracy. This allows for the design of Skill Scanner, an application that shows which skills are present in a CV or learning curriculum and how important they are to employers.

4.1 Importance of Skill Sets to Employers

The K-means model has learned 31 clusters from the dataset of 18.786 job postings. In order to evaluate the importance of each of these clusters to employers these requirements were taken back to the original dataset of job postings. The dataset of job postings was enriched with 31 columns: one column for each cluster. Then all skills in a job posting were evaluated and if a cluster was represented at least once in a given job posting, the corresponding column is 1. If a cluster is not present in a job posting, the corresponding column is 0. Averaging these columns determines the relative presence of each cluster in job postings. Figure 4.1 illustrates this technique.



Job	Requirements	C1	C2	C3	C4	C...	C31
Data Scientist	[Proficient in Python, Excellent stakeholder management, Automate ETL processes]	1	1	0	1	1	0
Managing Data Scientist	[Python skills like no other, etc. etc.]	1	0	0	1	0	0
Junior Data Scientist	etc.	1	0	1	0	1	1
Average		**1**	**0,33**	**0,33**	**0,67**	**0,67**	**0,33**

Figure 4.1: Calculation of Importance of Each Cluster to Employers.

4.2 Computing Coverage of Employers' Demand

The modelling pipeline infers clusters from it's input skills. The input skills come from a CV or the learning objectives of a study program. This way, a given input skill is labeled with the cluster of skills from the job requirement data that is most similar.

Finding the most similar skill set cluster is one objective, after finding this cluster, it is also required to find how well the skill covers it's inferred Skill Set cluster. To compute this, the cosine similarity of the input skill's embedding vector to the cluster centroid was used. For example, a cosine similarity of 0.7 is treated as a coverage of 70%.

4.3 An Estimation of Cluster Content

The K-means clustering has learned a set of arbitrary ellipsoids in a high dimensional feature space. These shapes in itself are meaningless but the members of each cluster are off course semantically similar. To communicate the meaning of a cluster to the end users, an estimation of each cluster's content is needed.

The data point closest to the cluster centroid could be a reasonable choice. However, allthough close to the centroid, this estimation might miss a large part of the cluster's semantic content. For example:

a requirements like "2 years of experience in programming in Python" is likely to be clustered together with requirements on programming in R. Choosing the data point closest to the cluster centroid would probably miss one of these programming languages and thus not communicate clearly what the cluster entails.

Another viable option to show the content of a cluster is the use of most frequent words. This technique showed unsatisfactory because many of the clusters content estimations would start with the word "data".

All though a bit of an outdated technique, in the past, bi-grams have been shown to be useful in similar contexts [Tan et al., 2002]. A set of three most frequent bigrams present in all data points of each cluster show to be quite informative. This technique was used to communicate the content of each cluster in all of the application's functions.

4.4 Skill Scanner Functions

The combination of thechniques as described in section 4.1, section 4.2 and 4.3 allows for an application that shows:

- Which skill clusters are present in an input skill set.
- How well each skill is covered in the input skill set.
- How important each Skill Cluster is to employers.

On these capabilities the three main functions of Skill Scanner were build:

4.4.1 Fit to Demand

Skill Scanner shows which skill sets are present in an input skill set and how important the skills are to employers. This functionality was named "Fit to Demand". To clearly communicate this information to end users the visual of which an example is shown in figure 4.2 was developed.

The content of each skill set is represented by the 3 most common bigrams in the job requirements dataset as explained in section 4.3.

The skill set importance was computed as described in section 4.1. For example, the "programming skills" cluster is present most often in job postings. This is why it is represented as the most important cluster.

The coverage of each skill set cluster was computed as described in section 4.2. For example, in the input skill set, a skill that covers the cluster "knowledge sql" was present in the input data, the input skill covers the skill set quite well.

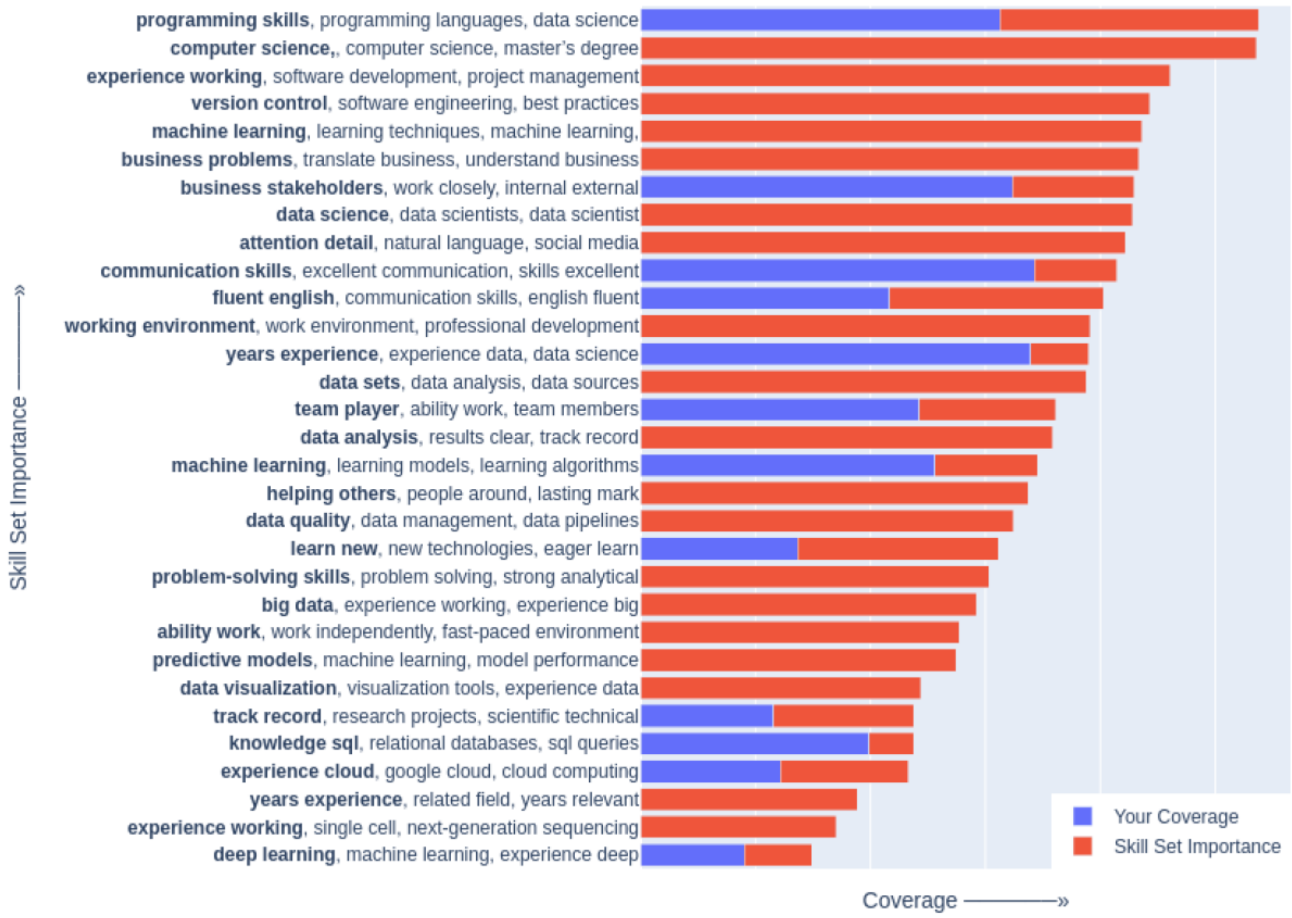


Figure 4.2: An example of Skill Scanner: Fit to Demand.

4.4.2 Comparison to Competition

To compare a skill set in a CV to a set of other CVs the Comparison to Competition functionality was developed. As it was shown in subsection 4.4.1, Skill Scanner computes how well each skill set cluster is covered. The arithmetic mean of this coverage is returned as the total coverage score.

To compare a given input CV to a set of competitors' CVs a binning approach was adopted. Skill Scanner computes the total coverage score of each input CV and bins them as shown in figure 4.3.

To create the bins for the example as shown in figure 4.3, the dataset described in section 3.3.3 was used.

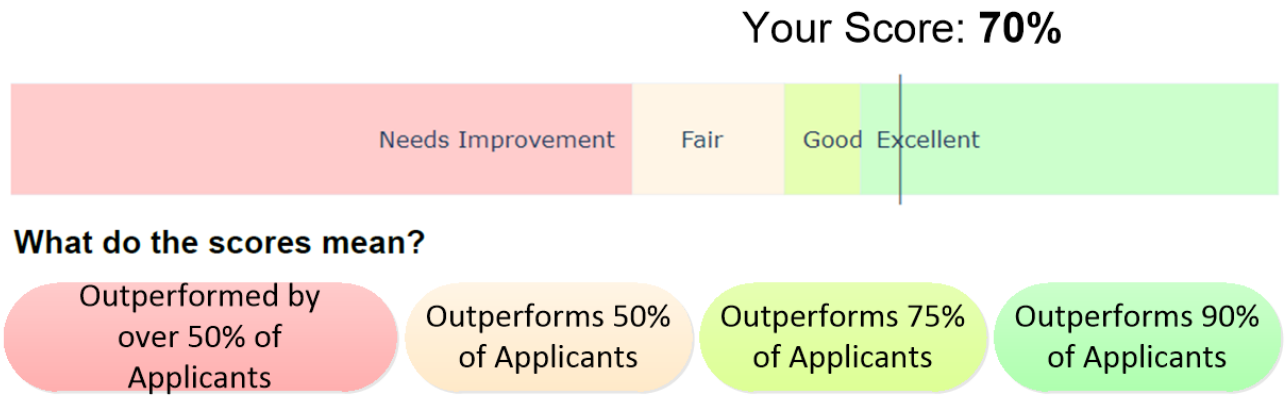

Your Score: 70%
Needs Improvement
Fair
Good
Excellent
What do the scores mean?
Outperformed by over 50% of Applicants
Outperforms 50% of Applicants
Outperforms 75% of Applicants
Outperforms 90% of Applicants

Figure 4.3: An example of Skill Scanner: Compare to Competition.

Please note: Because the centroid of a cluster in the feature space of sentence embedding vectors is meaningless, it is impossible to achieve a coverage of 100% for any skill. It is therefore important to communicate to end users what is a good score.

4.4.3 Find and Select

- Because Skill Scanner can compute which skills sets are present in a given set of input skills, it can also compute which skills are not present. These missing skills can be seen as skill gaps for an aspiring data scientist.

- Skill Scanner can analyze learning programs and find which skill sets are present in said program.

  Combining these points, Skill Scanner can recommend appropriate learning programs to enhance those skills that are:

  - Missing grom the input CV.
  - Present in the recommended learning program.
  - Important to employers.

  An example of such a training program recommendation is shown in figure 4.4. The recommendation shows which modules in the program "Msc Data Scientist 60ECT of IU International University of Applied Sciences" would be most appropriate for the author based on his CV.

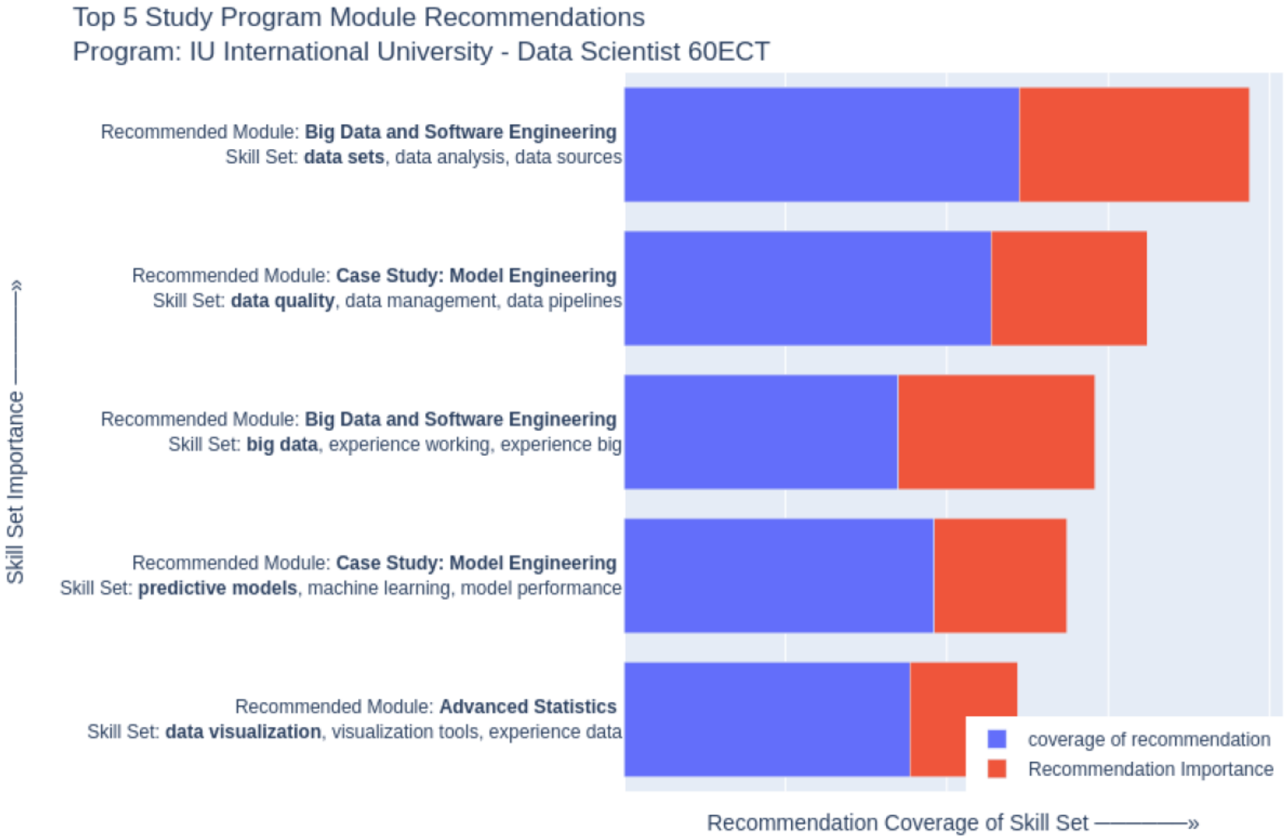


Figure 4.4: An example of Skill Scanner: Find and Select.

4.5 Relevance from Three Perspectives

The delivery of a beneficial product to all three parties involved has been one of the central ideas for this work. Consider figure 4.5. Skill Scanner revolves around supporting users from each of the three perspectives in their interaction with the other two perspective parties involved.

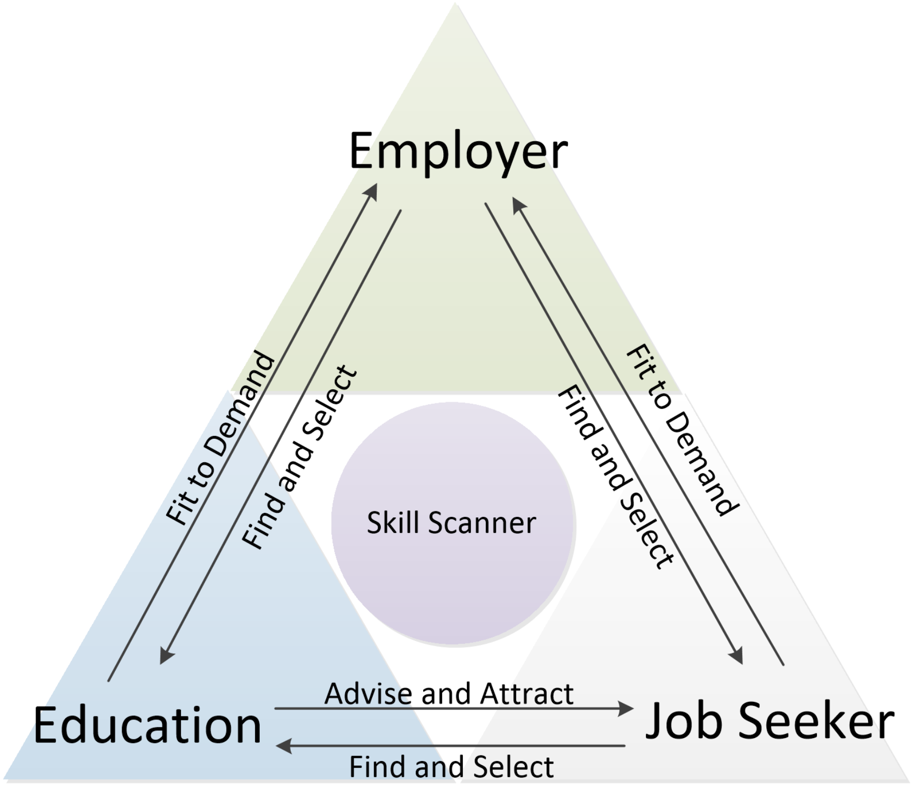


Figure 4.5: The Various Use Cases of Skill Scanner

4.5.1 The Employer's Perspective

- Find and Select: Skill Scanner provides insight in the skill gaps of employees. This makes it easier for employers to communicate with educational institutes and arrange appropriate schooling for their people.

  Skill Scanner can also help employers by recommending appropriate training programs based on the learning objectives of various training programs in comparison to the skill gaps of employees.

- Find and Select: Skill Scanner helps employers select the best fit for a job position by comparing a set of CVs to their vacancy's requirements. AI-based techniques for CV parsing and matching are not new. However, Skill Scanner does offer a novel approach that offers employers explainable decisions at the individual CV level. Skill Scanner could even be used to automatically generate reports for rejected CVs. This could potentially improve the candidate experience even when receiving a rejection.

4.5.2 The Job Seeker's Perspective

- Fit to Demand: Skill Scanner can help job seekers by showing which skills are present in their CV, how well they cover employers' demands and how important they are to employers. This can help job seekers decide what learning goals to focus on to make themselve more employable in field they aspire.

- Find and Select: As Skill Scanner can show which skills are present in a job seeker's CV, it can also identify which skills are lacking. The training program recommendation engine can then be used to recommend appropriate training programs to cover these skill gaps.

4.5.3 The Education Perspective

- Advise and Attract: In the same way Skill Scanner can recommend training programs to job seekers, it can also recommend training programs to students. By analyzing a potential student's CV, Skill Scanner can help determine the most appropriate learning program. Offering personalized and explainable advice can help education institutes in attracting and advising aspiring students.

- Fit to Demand: In a field evolving as quickly as Data Science, it is difficult for learning institutes to keep their learning programs relevant. Skill Scanner offers an automated and unbiased way to analyze the current job market. The application uncovers which skills are missing in a learning program and how important these skills are to employers. This can help educational institutes to periodically assess and adjust their learning content to better fit their programs to employers' demands.

4.6 Deployment

In order to showcase the capabilities of Skill Scanner a web app was developed which takes in a list of skills from a web-form. The skills are embedded and clustered using the methodology as described in chapter 3. The app then and returns a PDF with a set of reports based on the functionality as described in section 4.4. The app was containerized using Docker and was deployed as a Cloud Run instance on Google Cloud Platform. A functionality overview is available in figure 4.6. The application is available at [Bothmer, 2021b], the code and Docker container are avaiable at [Bothmer, 2021c].

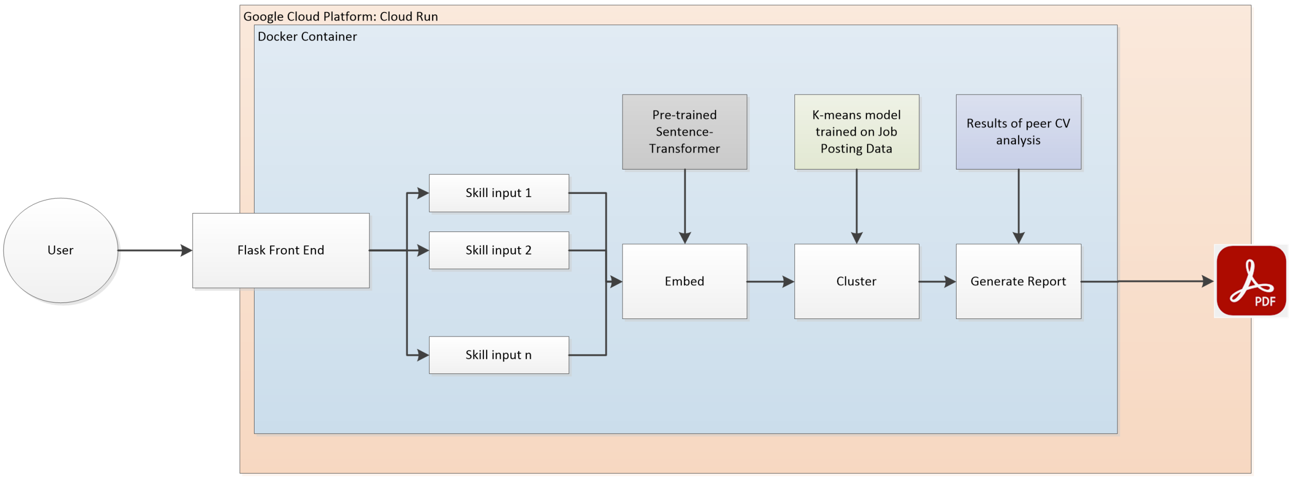

Google Cloud Platform: Cloud Run
Docker Container
Pre-trained Sentence-Transformer
K-means model trained on Job Posting Data
Results of peer CV analysis
User
Flask Front End
Skill input 1
Skill input 2
Skill input n
Embed
Cluster
Generate Report
PDF

Figure 4.6: Functionality Overview of Skill Scanner App

5 User Study

By the methods explained in chapter 3 and 4 the application "Skill Scanner" was developed and deployed. As described in chapter 4, Skill Scanner has many use cases for each of the involved perspectives. These use cases typically return reports: The user inputs a set of skills they want to compare to job market demands and the application returns a report adjusted to the user's needs.

As for this work, a proof of concept was developed which returns a set of reports based on the functions that are shown in section 4.4. An example of a report set is available in appendix **??**. These reports are aimed to be beneficial to a large array of potential user groups.

Because it is time and cost intensive to create reports more focused towards one specific user group it makes sense to seek feedback of potential end users. The importance of early user involvement in product development is an acknowledged priority [Kujala et al., 2002]. In the case of Skill Scanner, an early user study helps in answering questions like:

- Is Skill Scanner a useful product from the perspectives of potential end users?
- Would end users actually use Skill Scanner?
- What are the perceived benefits of Skill Scanner to potential end users?
- Which end user group would benefit most from a further developed version of Skill Scanner?

5.1 User Study Methodology

Figure 5.1 shows the structure of the user study's questionnaire. Potential end users were approached through e-mail and social media. All end users were asked to answer a set questions to acquire personal data like age, job role and gender. It makes no sense to ask use case questions to a user that is not familiar with a certain field. For example, it makes no sense to as an employer about the design of a learning curriculum. Therefore, the respondents were mapped to a specific part of the questionnaire that asked questions specific to their role. The mapping of respondents was done based on their answer to the question "Which point of view would best describe your PERSPECTIVE (if not any, just choose the one you can relate to most)?" After answering the perspective-specific questions, all potential end users were made familiar with the interface of Skill Scanner and provided feedback on it.

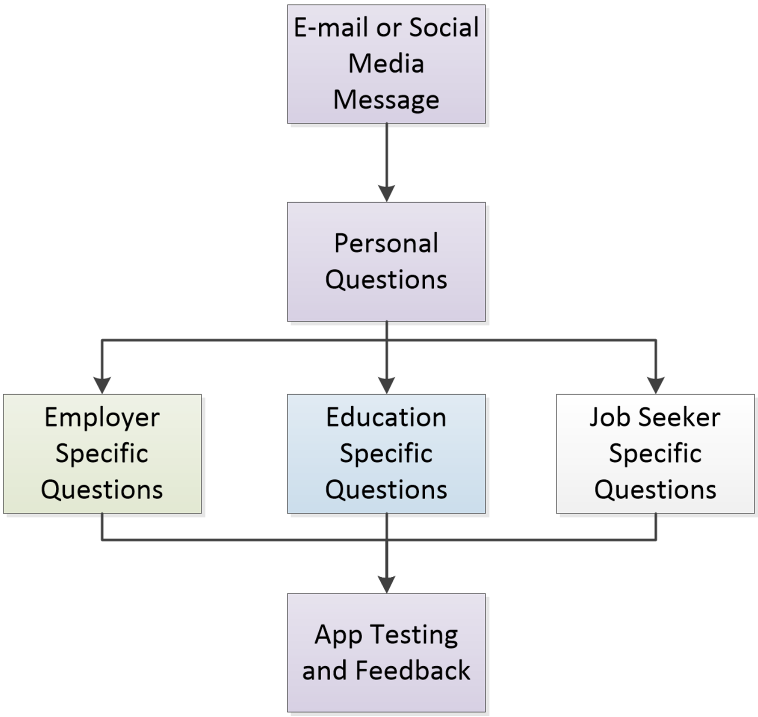
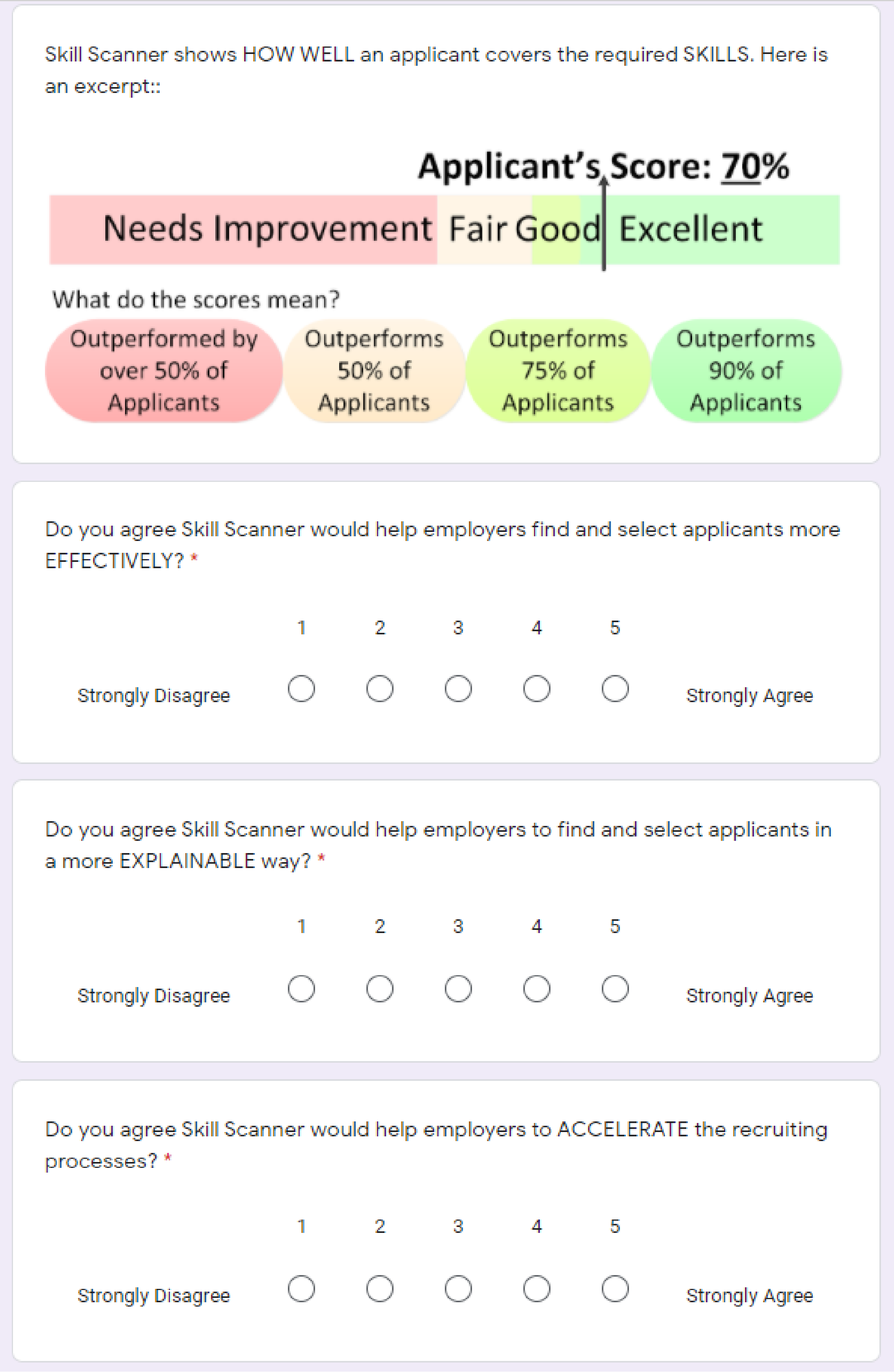

E-mail or Social Media Message
Personal Questions
Employer Specific Questions
Education Specific Questions
Job Seeker Specific Questions
App Testing and Feedback

Figure 5.1: The Structure of the Questionnaire.

Respondents provided feedback based on screenshots of Skill Scanner's functionality. It is debatable whether it would have been better to first let respondents use the deployed application before answering questions. Because long questionnaires are known to have a negative impact on response rate [Lund and Gram, 1998], the choice for screenshots was made in order to keep respondent's participation time short.

The respondents were asked a mixture of quantitative and qualitative questions. For quantitative results a set of 5-point scale likert-type question were asked [Boone and Boone, 2012]. An excerpt of the questionnaire is shown in figure 5.2.



Skill Scanner shows HOW WELL an applicant covers the required SKILLS. Here is an excerpt::

Applicant's Score: 70%
Needs Improvement
Fair
Good
Excellent
What do the scores mean?
Outperformed by over 50% of Applicants
Outperforms 50% of Applicants
Outperforms 75% of Applicants
Outperforms 90% of Applicants

Do you agree Skill Scanner would help employers find and select applicants more EFFECTIVELY? *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

Do you agree Skill Scanner would help employers to find and select applicants in a more EXPLAINABLE way? *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

Do you agree Skill Scanner would help employers to ACCELERATE the recruiting processes? *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

Figure 5.2: An Exceprt of the User Questionnaire.

The questionnaire was designed in English and translated to Dutch and German, the main languages in the countries where the authors of this work reside. This was done in order to encourage a larger target group to participate. The results were translated back to English in order to allow for a unified report.

5.2 Quantitative Results

5.2.1 Participants

75 participants were recruited (49.33% male, 50.66% female). Most respondents reside in either the Netherlands (60%) or Germany (37.33%), there were a few respondents residing in the UK (2.66%). The age distribution of participants is somewhat left skewed as can be observed in figure 5.3. The roles of the participants varied with teachers (29.33%) as the largest group in the sample. Figure 5.3 shows the distribution of roles reported by the participants. The participants are fairly well distributed over the three perspectives (29.33% Job Seeker, 32% Employer and 38.66% Education).

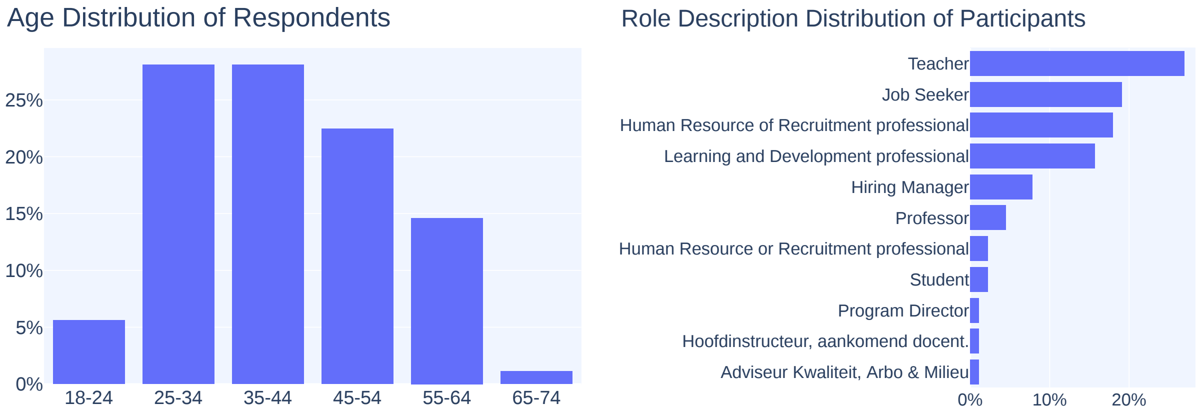


Figure 5.3: The Age and Role Distribution of Respondents.

5.2.2 Time Spent

Time is an important variable to understand which use cases would be most beneficial to end users. Figure 5.4 shows the estimated time respondents have spent on tasks which Skill Scanner could accelerate.

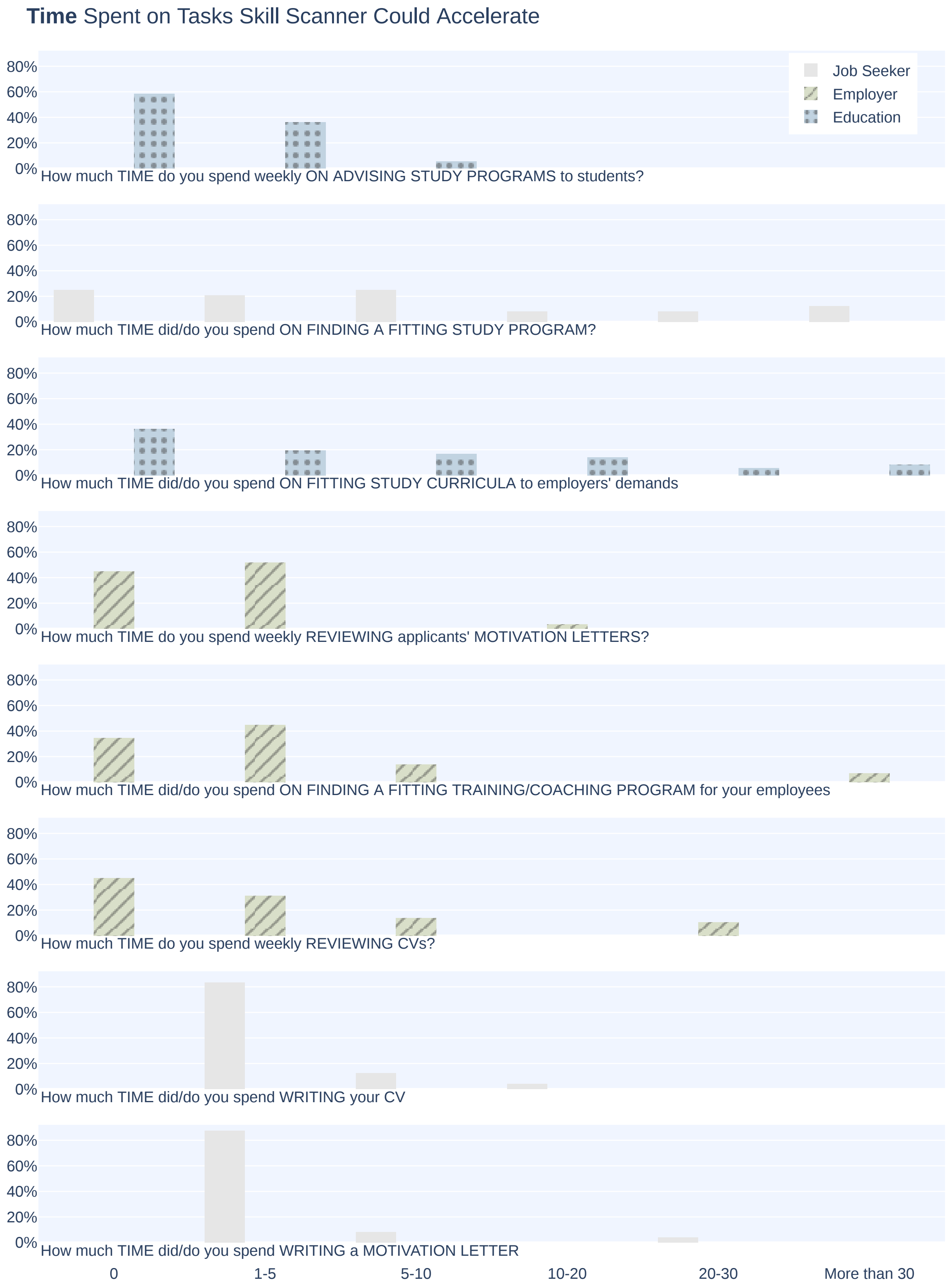


Figure 5.4: Amount of Time Respondents Spend on Various Tasks.

Discussion of Result Quite surprisingly, at least to the author, the time participants from all perspectives have spent on CV related tasks tends to be lower than the time spent on study program related tasks. Not only educational institutes spend a lot of time in these categories, employers and job seekers also tend to spend quite some time in finding a fitting study program. Where most of the efforts up until this point have been focused on CV optimization, it might be beneficial to focus Skill Scanner's development more towards study program selection and optimization.

5.2.3 Helpfulness of AI in general

Participants were asked questions on their perceived usefulness of AI in general for tasks from their perspective. In example, participants from each perspective answered a question on the usefulness of AI for "Finding and Selecting" from their perspective:

- Question for job seekers: Do you agree AI (Artificial Intelligence) based methods can HELP job seekers to FIND JOBS? (Optional)
- Question for employers: Do you agree AI (Artificial Intelligence) based methods can HELP with RECRUITMENT? (Optional)
- Question for education: Do you agree AI (Artificial Intelligence) based methods can HELP PLANNING AND REVISING a study program's learning curriculum? (Optional)

Figure 5.5 shows the summarized results for these questions.

Discussion of Result Overall, participants show quite high agreement about AI being helpful in relevant tasks. Job seekers returned the most positive outcome whereas employees of educational institutes were most sceptical about the helpfulness of AI in general. Possibly this effect is caused by the state of AI for each perspective. As described in chapter 2, the adoption of AI and data culture in education has been hesitant at best. A large proportion of the participants from education are teachers or even professors. Both highly educated professions who are trained to be critical. The agreeableness of job seekers was observed in all questions of the user study. This is perhaps related to job seekers being less deeply involved in the matters; a job seeker has a period of job seeking but applying it is usually not their core business for a prolonged period of time.

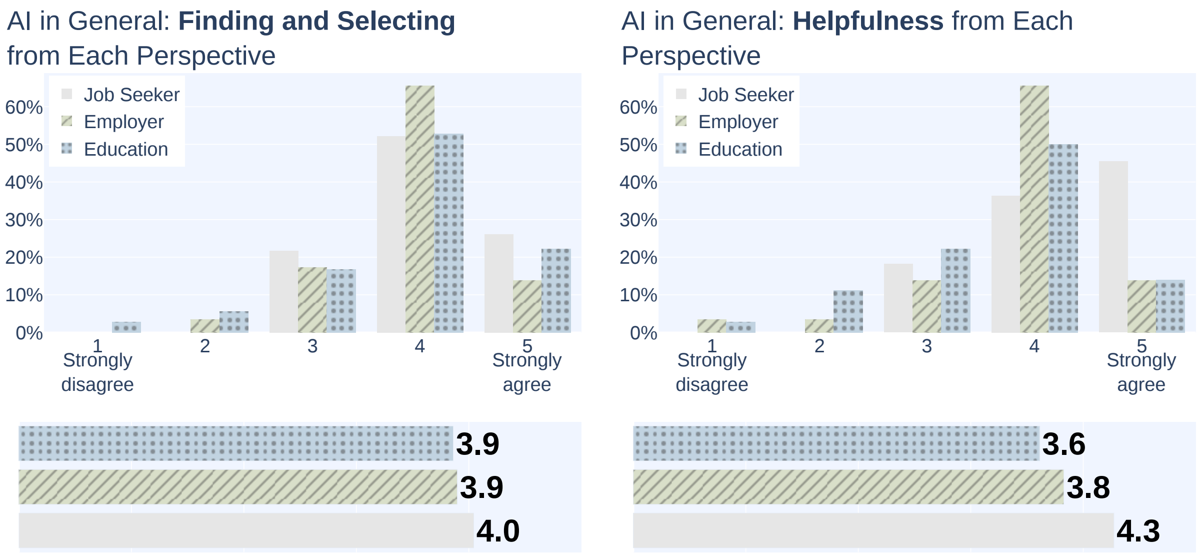


Figure 5.5: The Helpfulness of AI.

5.2.4 Benefits of Skill Scanner

Participants were asked to answer a range of 5-point Likert-schale questions on the potential benefits of Skill Scanner. For each question an equivalent suited to the perspective of the participant was asked. In example, the ability of Skill Scanner to accelerate processes as shown in figure 5.6 was measured with the following questions:

- Question for job seekers: Do you agree Skill Scanner would help job seekers to ACCELERATE their job seeking process?
- Question for employers: Do you agree Skill Scanner would help employers to ACCELERATE the recruiting processes?
- Question for education: Do you agree Skill Scanner would help learning institutes to ACCELERATE their study program design processes?

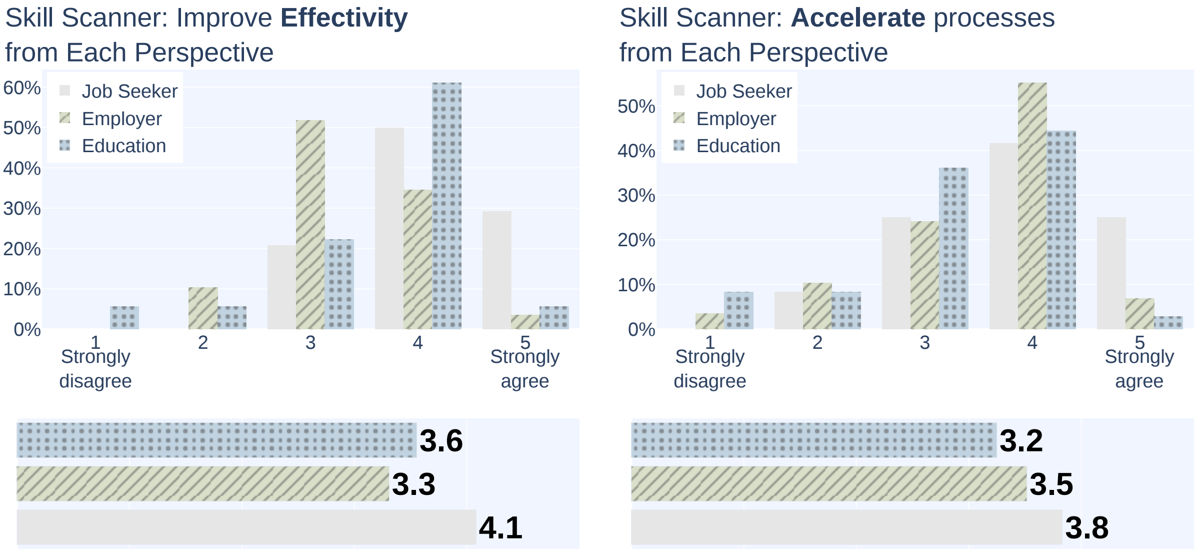


Figure 5.6: Accelerating and Improving Effectiveness of Processes Using Skill Scanner.

Discussion of Result: Effectivity The participants from the education perspective quite often agreed to the abilities of Skill Scanner to improve effectivity. The specific question they answered is:

> Do you agree Skill Scanner would help education institutes to design study programs more EFFECTIVELY?

Given the overall tendency of participants of the education perspective to be the least agreeable, this result seems to be a strong indication that there lie opportunities to expand on Skill Scanner's functions to better facilitate education institutes in designing their study programs.

Discussion of Result: Accelerate Aside the tendency of job seekers to be positive, participants were neutral to somewhat agreeable about the abilities of Skill Scanner to accelerate processes. The least agreeable to this question were participants from the employers perspective. This might be caused by other CV parsing products with which Skill Scanner would have to compete; the use of AI in recruitment practices is already adopted making it more difficult to accelerate these processes. This might indicate that the gains in supporting employers in finding and selecting fitting employees are marginal because of the existing competition.

For the same reasons, job seekers might tend to agree so much more; As it is unlikely for job seekers to have actively used AI in their job seeking processes, the opportunities tot accelerate these processes are relatively large.

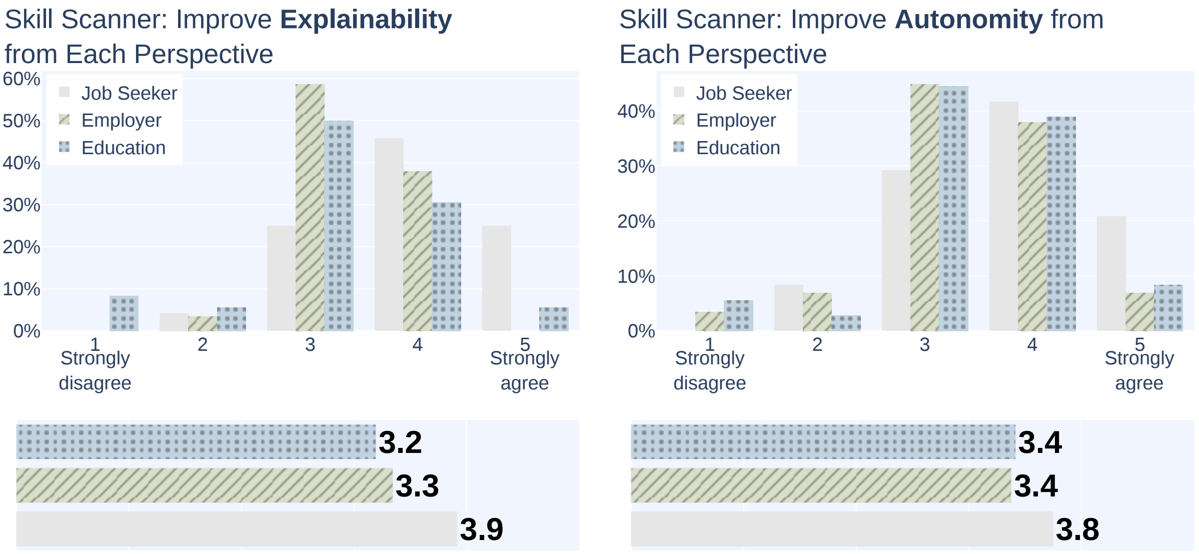


Figure 5.7: The Perceived Abilities of Skill Scanner to Improve Explainability and Autonomity.

Discussion of Result: Explainability Participants are neutral to somewhat agreeable about the abilities of Skill Scanner to improve explainability. Employers remained relatively neutral on the matter whereas participants from the education perspective were more outspoken; they agreed or disagreed more strongly. Job seekers were the most agreeable but the score is relatively low considering job seekers' answers to the other questions.

Discussion of Result: Autonomity The benefit of Skill Scanner for improving autonomity of learning processes requires quite a stretch of the imagination. Therefore it was quite surprising to the author that participants were quite agreeable on this question. All tough this benefit does not really offer opportunities to expand Skill Scanner's functions, it does hint to a unique selling point of the product. Just like the explainability question, job seeker' agreeableness is high absolutely, but relatively low with respect to job seekers' answers to other questions.

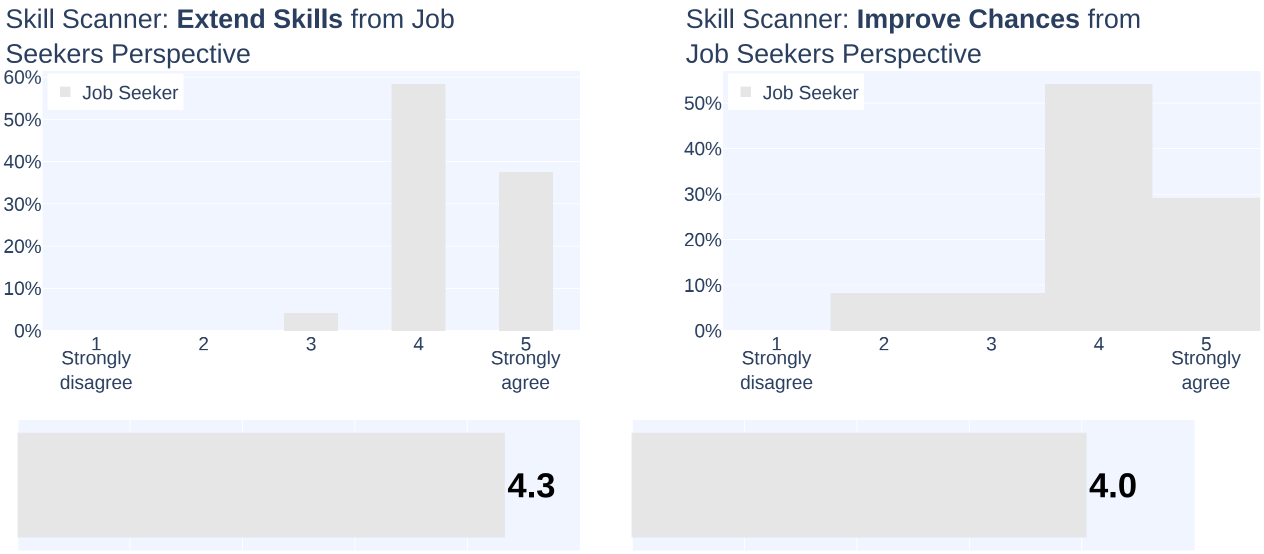


Figure 5.8: Extending Skills and Improving Chances on Job Market of Job Seekers

Discussion of Result: Extend Skills With an average score of 4.3, this is the strongest agreeableness observed in all questions to job seekers. This result might be explained by the fact that Skill Scanner's functions, as demonstrated to participants, are optimized for this task. The main function of Skill Scanner is to show which skills are present in a CV, which skills are missing, and how important they are to employers. In addition, it shows were to learn these specific skills.

Discussion of Result: Improve Chances Job seekers agreed quite strongly about Skill Scanner's abilities to help job seekers improve their chances of finding a fitting job. This is an important question as it describes job seeker's end goal of using Skill Scanner in the first place.

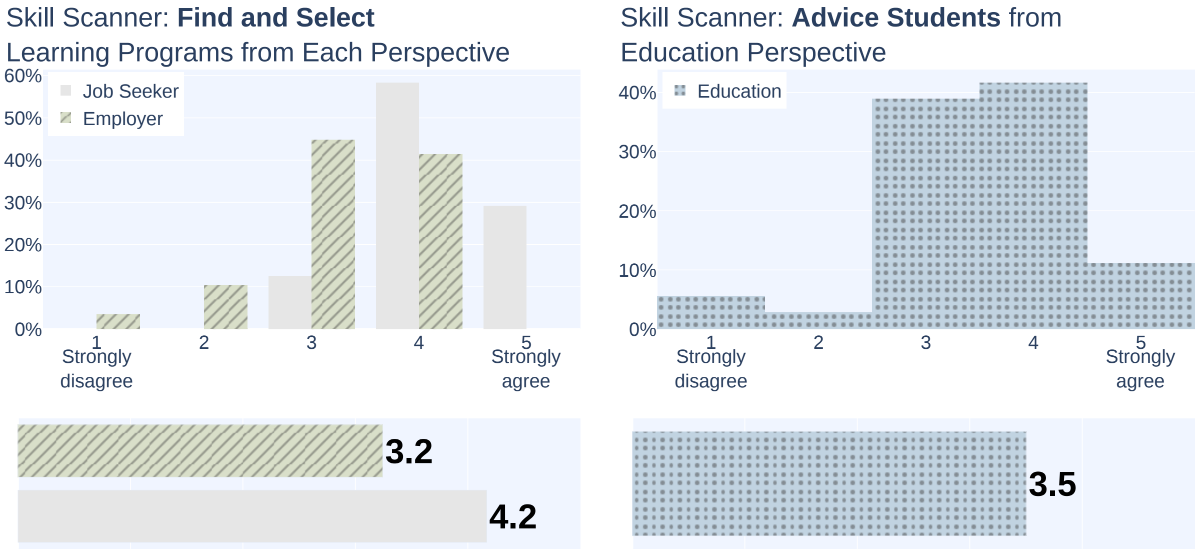


Figure 5.9: The Helpfulness of Skill Scanner to Find and Select Learning Programs and Advice Students.

Discussion of Result: Find and Select Learning Programs Were job seekers were quite agreeable to the abilities of Skill Scanner to help them find and select a fitting study program, employers remained more or less neutral on the matter. While job seekers seem biased to respond somewhat more agreeable, there might be a second cause of this difference: a job seekers can just upload a CV to Skill Scanner and get study program recommendations. For an employer, the use case is less clear, they do not really need an analysis of an employees CV to determine what learning goals they should focus on.

Discussion of Result: Advising Students Participants from the education perspective were pretty agreeable about Skill Scanner's abilities to help advice appropriate study programs to their students. This use case leverages the full capabilities of Skill Scanner to compare skills from three sources:

1. Learn what is important to employers from job postings.
2. Analyze which skills are present and which skills are missing from a student's skill set or CV.
3. Recommend an appropriate study program based on content of learning programs.

The answer to this question shows participants from the educational perspective perceived this functionality as helpful, even though the functionality is now optimized for job seekers. This result is an indication that it might be fruitful to invest in Skill Scanner's abilities to help in advising study programs from an educational perspective.

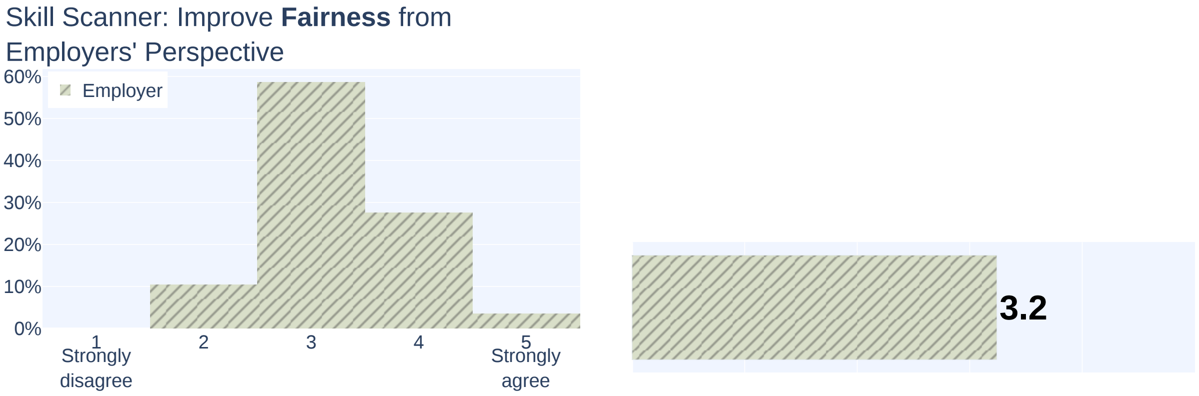


Figure 5.10: The helpfulness of Skill Scanner for fair assessment in recruitment.

Discussion of Result Employers are more or less neutral about Skill Scanner's abilities to improve fairness. There are not many opportunities to improve or better communicate this. It seems like improving fairness is not something to expand on neither a unique selling point of Skill Scanner.

5.2.5 Usefulness of Skill Scanner

In order to measure the perceived usefulness of Skill Scanner, participants were asked whether they would use Skill Scanner for the typical tasks from their perspective:

- Question for job seekers: Would you USE Skill Scanner in your job seeking process?
- Question for employers: Would you USE Skill Scanner in your work?
- Question for education: Would you USE Skill Scanner in your work?

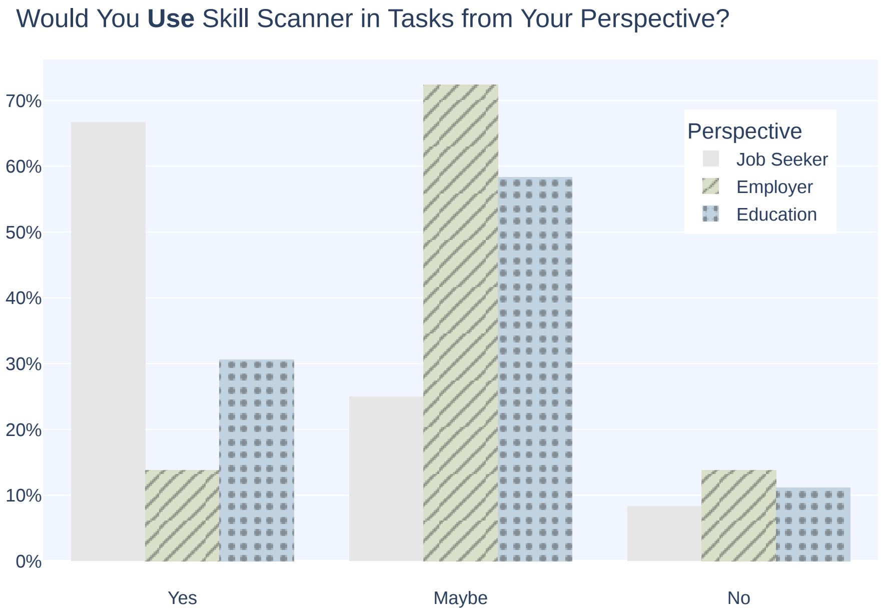


Figure 5.11: Participant Would Use Skill Scanner from Each User Perspective.

Discussion of Result Most participants were willing to use or maybe use Skill Scanner in their processes. Job seekers were the most willing to use skill scanner where from the education and employers' perspective the mode was maybe. This might be explained by how easy it is for job seekers to use Skill Scanner. For the other two perspectives adopting Skill Scanner would mean a significant change in the way they work. It is explainable that this causes these participants to be more hesitant in their answer to this question. Overall it is a positive result that participants were quite unlikely to answer "No".

5.3 Qualitative Results

Aside the quantitative questions, participants were asked for additional feedback and remarks in order to allow for more open feedback then a Likert scale. While these results are more challenging to report as inclusively as quantitative questions, this summary highlights the main topics mentioned by the participants. Some quotes were reported that are most illustrative for the remarks on each topic. For the full set of answers, it is are fully available at the Github repository supporting this thesis [Bothmer, 2021c]. In general, 12% of all participants answered the question "Do you have any REMARKS regarding using Skill Scanner?". 8% answered the question "Do you have any REMARKS regarding Skill Scanner?". 18% provided an answer to the question "Do you have any ideas for IMPROVEMENT or use cases of Skill Scanner?". Results from the Dutch or German versions of the questionnaire were translated to English.

Whilst different questions, the overarching themes in the answers are similar. The result summary is provided per theme.

5.3.1 Soft Skills

Many participants mention that Skill Scanner, and AI in recruitment practices in general, focuses on the assessment of hard skills. Some respondents mention the inclusion of soft skills as a possible future improvement for Skill Scanner. One specific soft skill that participants mentioned is motivation. One participant described her considerations regarding this topic:

> "Skill scanner in my opinion only assesses hard skills. In a time where soft skills are becoming more and more important, I think this is a missed opportunity. You can teach someone hard skills. Developing soft skills is more difficult. Selecting on soft skills should be more prominent in selection processes."

5.3.2 Human in the Loop

When asked for remarks on Skill Scanner, many participants expressed doubts about assessment of humans by machines. Some argue an application like Skill Scanner could support employees and employers to have more meaningful dialogue. Others doubt AI even has a role in this part of the process. Two participants' quotes best describe these reflections:

> There should rather be templates for exchange, listening, joint reflection processes, etc. There should be scanners that understand the relationship between employee and employer as a professional but still personal exchange process, an approach from both sides - and not as a checklist.

> With an applicant I also pay attention to whether he fits into the team, his appearance, motivation and enthusiasm. Skill Scanner can provide a scientific vision of an applicant. The human side, certainly also important, remains human work.

5.3.3 Supporting Role of AI

Some participants note that the role of AI should remain supportive and not prescriptive. An example quote of one of our participants:

> It should remain a tool of assistance or guidance and not an all encompassing thing.

5.3.4 Practical Considerations

Participants added a lot of tips to improve Skill Scanner. Ranging from validating skill inputs with short tests to asking more questions to make the experience more personal. Some of these results are more related to the interface of Skill Scanner as a product. These are described in section 5.4.

5.3.5 Discussion of Qualitative Results

The open questions provide quite some insight as to why people are hesitant to adopt AI based products like Skill Scanner. The underlying sentiment among the reactions of most of the participants seems to be a reluctance to systems were machines assess humans. Participants strongly advice against functionalities like Skill Scanner's "Compare to Competition" as shown in figure 4.3 because it could be used to cut out human-to-human interaction.

The vision behind Skill Scanner is to improve interaction among job seekers, employers and educational institutes by providing consensus on what skills are most important to employers. The resulting product should improve the quality of interaction among these parties.

It is important to consider this sentiment and strongly communicate the focus of Skill Scanner on improving interaction.

5.4 Feedback on Interface

Whilst not a priority in this work, an application interface had to be developed. Participants were asked to provide feedback on the application interface. This part of the questionnaire consisted of one quantitative question followed by four open questions to allow participants to express their ideas, sentiments and suggestions on Skill Scanner's interface. The results of the quantitative question are shown in figure 5.12.

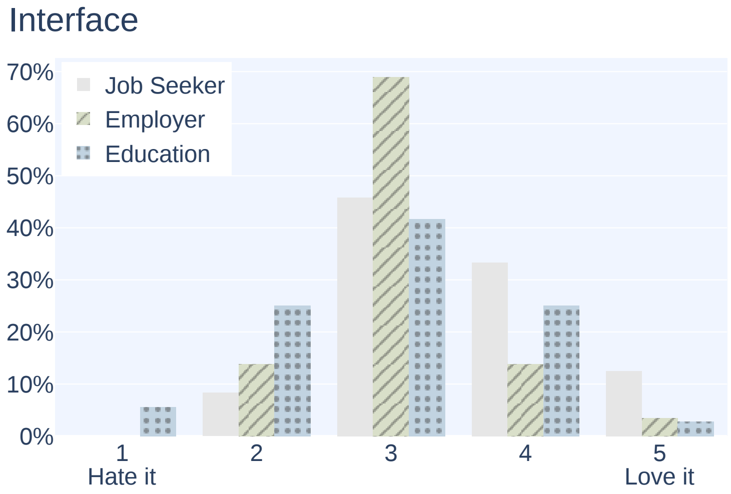


Figure 5.12: How Participants Like Skill Scanner's Interface.

5.4.1 Qualitative Feedback on Interface

Participants were asked four questions. "What is the main reason for your score?" was answered by 72% of participants. "What would you CHANGE in Skill Scanner's interface?" had a response rate of 58%. "What would you ADD to Skill Scanner's interface" had a response rate of 39%.

There are a lot of tips and tricks mentioned in the answers. The full set of answers is available at the Github repository supporting this work [Bothmer, 2021c]. The general consensus on the interface is that it is clear and consise but rather basic and plain. Many participants mention it is not visually pleasing or 'sexy'. Respondents mention the use of color and fonts could be improved. Some quotes that illustrate these reflections:

> A bit basic, you may be a bit bolder with colors. Look for active and motivating colors on the internet.

> The score is based purely on the look and feel, looks fine but nothing special.

5.4.2 Discussion of Feedback on Interface

The application interface was perceived as quite neutral. While the development of the interface was not a priority for this work, it is still important to acquire user feedback from as early as possible. The acquired scores are also useful as a benchmark to validate future improvements. The acquired feedback was carefully documented as these will be very beneficial in any future iteration of the product design cycle.

6 Conclusion and Discussion

6.1 Discussion of Methodology

The developed methodology consists of quite a long pipeline ranging from data acquisition to the developed modelling techniques as can be observed in figure 6.1. All design choices were made with care as described in chapter 3. Still, each of these steps offer their own opportunities to expand or improve the methodology. Therefore each step of the pipeline is discussed in this section.

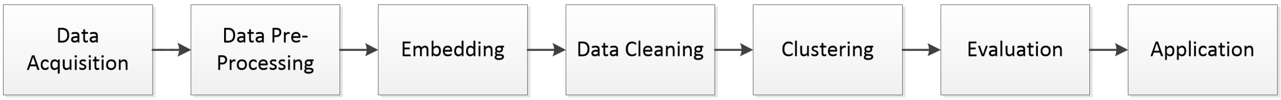


Figure 6.1: Overview of the Methodology.

6.1.1 Data Acquisition

As required skills are ever changing, it is important to keep the source data up to date. In this work, a moderately large dataset of job postings was acquired from various sources. In order to scale to more job titles and keep the models up to date, further automation of data acquisition is needed. This poses a challenge as most job posting websites employ active measures against web scraping. Web scraping is also against the terms of service of some of the large job posting websites. It is considered a legal grey area. All though Web scrapers seem to be on the winning hand [Sobel, 2020], products that are heavily relying on web scraping techniques will remain at risk as this legal discussion is far from over. A collaboration with one of the large job posting websites could potentially solve these issues.

6.1.2 Data Pre-Processing

The methods to extract skills from data sources in this work are rather basic, relying on simple text splitting techniques for skill extraction. These techniques showed to be remarkably effective all though it also caused a large loss of data points from the job postings dataset. For future work, it is recommended to adopt more sophisticated skill extraction techniques. It must be noted that it is important that any improvement in this technique should remain automatable. Automatic skill extraction is important to keep products build on the methodology scalable.

6.1.3 Embedding

The Sentence-Transformers library [Reimers and Gurevych, 2020] is fundamental to this work. The pre-trained general purpose models all performed remarkably well on the extracted datasets. In future improvements it could be beneficial to the accuracy of the model to train a specific Sentence-Transformer model based on a large dataset of data comparable to Skill Scanner's input data. The datasets acquired for this work proved to be too small for such an effort.

6.1.4 Data Cleaning

The data cleaning step using a combination of UMAP and DBSCAN was very advantageous to the quality of the modelling done afterwards. This step is an attention point should the methodology be generalized to other job titles; The tuning of DBSCAN's hyper-parameters took some manual iterations. It will be interesting to see if these hyper-parameters are correct for cleaning data from different job titles.

6.1.5 Clustering

While K-means clustering is a proven and sensible method for the use case of this thesis, it certainly is not the only candidate for the task. It could make sense to cluster a projection of the embedding vectors by using a strong dimensionality reduction technique. UMAP to two dimensions yielded very good results, all though the K-means models on the full dimensionality still outperformed these attempts. Experimenting with the number of dimensions for UMAP to reduce to could result in a projection that enables strong clustering candidates that might outperform the developed K-means model.

6.1.6 Evaluation

The developed K-means model generalized to data from different sources with reasonable accuracy (77% for learning curricula and 88% for CV data). There is certainly room for improvement here. One recommendation is to introduce an outlier removal step to the processing of data from learning curricula and CVs. In the work as described, the model always has to infer a cluster. A large part of the miss-classifications could be considered outliers. A method similar to the data cleaning of the job postings could be developed to automatically clean input data before inferring the clusters.

6.2 Discussion of Application

Because this thesis proposes a novel product it is rather challenging to interpret the results. There is no benchmark product to which the outcomes can be compared. As described in chapter 1, this work describes the first iteration of a product design feedback loop. The results of the user study can be used as input to benchmark future improvements to the product.

There is a lot of room for improvement to the developed application. The interface should be improved, the user input from chapter 5 will be helpful for this. It should be possible to upload a CV as a file instead of manually putting in a set of skills.

One of the key elements of this work is the synthesis of skills from three perspectives. The ability of the model to infer skill clusters of data from these three sources makes it possible to compare inherently differently formulated skills. This gives our application Skill Scanner a very strong base on which all kinds of product layers can be build. The created reports are just samples to show the capabilities of Skill Scanner. In a future iteration the reports should be more tailored to each perspective and every user should receive a further personalized report. For each perspective some recommendations are made below:

6.2.1 The Education Perspective

The user study as described in chapter 5 shows evidence that participants from the education perspective agreed to the ability of Skill Scanner to improve effectiveness in study program design. As was demonstrated in chapter 4, the "Fit to Demand" functionality of Skill Scanner provides a good base for such use cases. It shows which skills are present in a learning program, which skills are missing and how important these skills are to employers.

A future improvement could be to acquire a large amount of CVs. These could be used to compare the skills from CVs that show continuous growth to those that do not. In this way study programs can be designed that set students up for future success, giving them a real edge after they graduate. This is an interesting project in it's own for which the learnings from this work could be very beneficial.

Participants from the Education perspective were also agreeable about Skill Scanner's helpfulness in advising their students. Their responses to these categories were considerably more positive indicating potential use cases to expand on in future work. This result was yielded by explaining Skill Scanner's "Find and Select" function to education participants. As explained in chapter 4, "Find and Select" shows which skills a potential student needs to learn based how important they are to employers. From this analysis the app recommends study programs to learn these skills.

A recommendation to expand this functionality could be to add a job role recommendation to Skill Scanner. Based on skills a potential student already possesses, Skill Scanner could look for:

- CVs in which these skills are also present. From this data it could recommend common job roles among these CVs and also common supplemental skills among these CVs.
- Jobs that require these skills. From this data it could recommend these jobs and which supplemental skills to learn.

Adding these recommendations would unlock Skill Scanner's full capabilities to support educational institutes in their processes.

6.2.2 The Employer's Perspective

The responses of participants from the employers perspective were considerably more neutral and homogeneous. This makes it difficult to determine were to focus future efforts on applications for the employers perspective.

The "Compare to Competition" functionality demonstrated to employers is not as novel as some other functions of Skill Scanner.

The "Find and Select" functionality that was shown to employers in the user study would help employers to find fitting study programs for their employees, based on the CV of said employee. This requires quite a stretch of the imagination and the use case is perhaps to niche to be worthwhile for a large audience.

While Skill Scanner still shows potential for being beneficial to employers, the capabilities Skill Scanner has to offer to them are not as unique as for the education perspective. This might explain the neutral responses. For future work the use cases for employers are not the most sensible candidates to further develop.

6.2.3 The Job Seeker's Perspective

The most agreeable participants showed to be job seekers, who quite strongly agreed to all potential benefits of Skill Scanner. This rises the question: were job seekers less critical in their responses or is this the perspective that offers most opportunities for Skill Scanner to expand on? Some qualitative research is recommended to further elaborate on this topic. A good next step could be to allow a range of job seekers to fully test Skill Scanner's functionality for them and interview them afterwards.

All job seekers agreed to all perceived benefits, one result stood out: The ability of Skill Scanner to help job seekers extend their skill set.

6.2.4 Synthesis

The user study provided a set of strong arguments to reposition Skill Scanner as a learning and coaching tool:

- Participants spend most time on learning related tasks as discussed in section 5.2.2.
- Participants from both the educational and job seeker's perspective reacted most agreeable to Skill Scanner's abilities to extend or advice skills.
- In the qualitative part of the user research, participants expressed quite strongly the need for a human in the loop and that the role of AI should remain supportive.

With some modifications, Skill Scanner could be turned into a product study advisors and job coaches can use to better support students or job seekers.

6.3 Conclusion

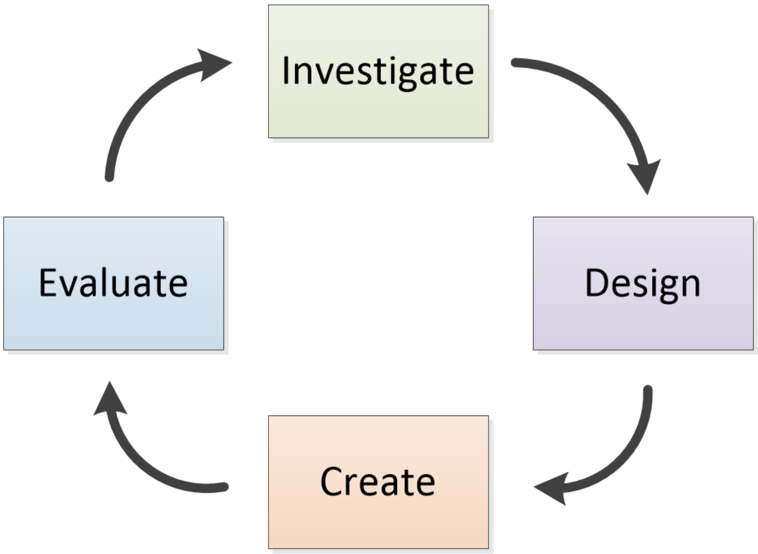


Figure 6.2: The Feedback Loop Analogous to the Thesis.

As described in chapter 1, this thesis is best understood as an iteration of a product design feedback loop. Figure 6.2 shows this idea. Each element of this loop was characterized by it's corresponding research question. Each of these research questions will be answered in the subsequent paragraphs:

Research Question 1: How is AI currently being used to improve interaction among job seekers, employers and education? The field of AI in human resource management and recruitment proved to be a scattered field. There are lots of proofs of concepts but a set of best practices seems non-existent. Companies seem hesitant to adopt data culture and AI in their human resource management and recruitment practices.

Extensive literature research yielded no results of synthesis efforts combining skill analysis from all job seeker's, employer's and education's perspectives in one product or analysis. Therefore, the literature of AI for skill extraction and analysis for each of these three perspectives was reviewed in isolation.

For job seekers, the field revolves around job recommendation and learning content recommendation. While an active research field in the past, the field seems to have matured and adopted by large tech companies like for example LinkedIn [Guo et al., 2019]. They are leveraging AI in these use cases to greatly enhance their products.

For Employers, the most important use case is the automated comparison of CV's to job requirements. Most of the related work on extracting skills from CVs is based on classification methods, e.g.: [Faliagka et al., 2014] [Fernández-Reyes and Shinde, 2019] [Celik, 2016] [Nigam et al., 2020]. These works offer supervised learning approaches to achieve automated CV comparison, the performance increases with the advancement of the natural language processing field. An unsupervised learning approach could expand on these methods by offering methodology that is easier to maintain from an MLops perspective.

For education, quite some efforts for recommendation have been proposed [Rivera et al., 2018]. It remains unclear how well they have been adopted as educational institutes seem very hesitant to adopt AI-based technologies [Luckin and Cukurova, 2019]. There are opportunities to expand on existing methodology by introducing data mining and artificial intelligence as concluded by [Rivera et al., 2018].

Overall, this thesis argues that the three parties involved could benefit greatly from having a consensus on which skills are important for a data scientist. This could improve interaction among all three perspectives involved.

Research Question 2: What are the most commonly required skills for a data scientist from the analysis of job postings? Can these skills be modeled so that they can be compared to skills extracted from different sources? A dataset of 2.633 Data Scientist job postings was acquired. From this dataset 21.509 job requirements were extracted. For these job requirements, sentence embedding vectors were attained using various state-of-the-art BERT-based pre-trained sentence embedding models [Reimers and Gurevych, 2020]. The data was cleaned using an innovative density based clustering approach. The remaining 18.786 job requirements were used to train a range of k-means clustering models. Using a grid-search with silhouette score as the criterion, the optimal model was selected. Some elementary arithmetic calculations were used to compute which skill clusters are most important to employers.

The model showed to generalize reasonably well to unseen data from different sources: The model inferred the correct skill cluster for 77% of a set of learning objectives from a sample learning curriculum. The model inferred the correct cluster for 88% of all skills described in a representative sample set of Data Scientist CVs.

This ability of the developed model to generalize to unseen data from different sources makes it possible to: Show which skill sets are present in a CV or learning curriculum and how important they are to employers.

Research Question 3: How can a skill model be applied as a service and be beneficial to job seekers, employers and educational institutes? Based on the capabilities of the developed methodology the application Skill Scanner was developed. Skill Scanner is a Flask-based web application that is deployed to Google Cloud Platform as a Docker container [Grinberg, 2018]. The application takes a set of skills as it's input and returns a set of reports that are tailored to the demands of each of the perspectives involved. The reports are based on the three main functions of Skill Scanner:

- Fit to Demand, Skill Scanner's ability to show which skills are present in a CV or learning curriculum and how important they are to employers.
- Comparison to Competition, Skill Scanner's functionality that compares a CV to a group of CVs with respect to employers' demands.
- Find and Select, the learning content recommendation engine of Skill Scanner.

The generated reports have great potential benefits to each of the perspectives involved.

Research Question 4: What are the strengths and weaknesses of the developed application? A user study based on a questionnaire took place. Overall, participants reacted fairly positive to the potential benefits of skill scanners to aid them in their tasks. Job seekers were most agreeable to the helpfulness of Skill Scanner. Participants from the education end employer's perspective proved to be more difficult to convince, responding more neutral to some questions. Participants from the education perspective showed particularly positive reactions on the capabilities of Skill Scanner to design study programs more effectively and the helpfulness of Skill Scanner in advising their students.

6.4 The End of a Beginning

With the answer of the last research question, considering the feedback loop as illustrated in figure 6.2, this iteration has come to an end. The keen observer will notice the arrow pointing back to the initial phase: Investigate. It is time to finish this work. Not as an end but as a new beginning. It is time to return to the first phase of the cycle for a second iteration, to *investigate* the ideas learned from chapter 5. From here new methodology can be *designed* to *create* an improved application and *evaluate* whether it is truly better.

Here is hoping that this second iteration will not be a singular effort by the author, and that that the ideas explained in this thesis will provide inspiration to others to synthesize the developed methodologies with their own ideas. As a firm believer in open science, all code was made available in a reproducible way on the Github repository supporting this thesis. This is unique in this line of work.

Bibliography

[Al-Otaibi and Ykhlef, 2012] Al-Otaibi, S. T. and Ykhlef, M. (2012). A survey of job recommender systems. *International Journal of Physical Sciences*, 7(29):5127–5142.

[Alammar, 2021] Alammar, J. (2021). The illustrated word2vec.

[Allaoui et al., 2020] Allaoui, M., Kherfi, M. L., and Cheriet, A. (2020). Considerably improving clustering algorithms using umap dimensionality reduction technique: A comparative study. In *International Conference on Image and Signal Processing*, pages 317–325. Springer.

[Angelov, 2020] Angelov, D. (2020). Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*.

[Baškarada and Koronios, 2017] Baškarada, S. and Koronios, A. (2017). Unicorn data scientist: the rarest of breeds. *Program*.

[Bennett, 2002] Bennett, R. (2002). Employers' demands for personal transferable skills in graduates: A content analysis of 1000 job advertisements and an associated empirical study. *Journal of Vocational Education and training*, 54(4):457–476.

[Boettiger, 2015] Boettiger, C. (2015). An introduction to docker for reproducible research. *ACM SIGOPS Operating Systems Review*, 49(1):71–79.

[Boone and Boone, 2012] Boone, H. N. and Boone, D. A. (2012). Analyzing likert data. *Journal of extension*, 50(2):1–5.

[Bothmer, 2021a] Bothmer, K. (2021a). Job analyis. `https://github.com/KoenBothmer/job_analysis`.

[Bothmer, 2021b] Bothmer, K. (2021b). Skill scanner application. https://skillscanner-x3gfl3cnea-ew.a.run.app.

[Bothmer, 2021c] Bothmer, K. (2021c). Thesis experiments. `https://github.com/KoenBothmer/Thesis`.

[Celik, 2016] Celik, D. (2016). Towards a semantic-based information extraction system for matching résumés to job openings. *Turkish Journal of Electrical Engineering & Computer Sciences*, 24(1):141–159.

[Cox and King, 2006] Cox, S. and King, D. (2006). Skill sets: an approach to embed employability in course design. *Education+ Training*.

[Davenport and Patil, 2012] Davenport, T. H. and Patil, D. (2012). Data scientist. *Harvard business review*, 90(5):70–76.

[De Mauro et al., 2016] De Mauro, A., Greco, M., Grimaldi, M., and Nobili, G. (2016). Beyond data scientists: a review of big data skills and job families. *Proceedings of IFKAD*, pages 1844–1857.

[Debortoli et al., 2014] Debortoli, S., Müller, O., and vom Brocke, J. (2014). Comparing business intelligence and big data skills. *Business & Information Systems Engineering*, 6(5):289–300.

[Decorte et al., 2021] Decorte, J.-J., Van Hautte, J., Demeester, T., and Develder, C. (2021). Jobbert: Understanding job titles through skills. *arXiv preprint arXiv:2109.09605*.

[Devlin et al., 2018] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[Diaby et al., 2013] Diaby, M., Viennet, E., and Launay, T. (2013). Toward the next generation of recruitment tools: an online social network-based job recommender system. In *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, pages 821–828. IEEE.

[Ester et al., 1996] Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231.

[Faliagka et al., 2014] Faliagka, E., Iliadis, L., Karydis, I., Rigou, M., Sioutas, S., Tsakalidis, A., and Tzimas, G. (2014). On-line consistent ranking on e-recruitment: seeking the truth behind a well-formed cv. *Artificial Intelligence Review*, 42(3):515–528.

[Fernández-Reyes and Shinde, 2019] Fernández-Reyes, F. C. and Shinde, S. (2019). Cv retrieval system based on job description matching using hybrid word embeddings. *Computer Speech & Language*, 56:73–79.

[Garg et al., 2018] Garg, V., Srivastav, S., and Gupta, A. (2018). Application of artificial intelligence for sustaining green human resource management. In *2018 International Conference on Automation and Computational Engineering (ICACE)*, pages 113–116. IEEE.

[Grinberg, 2018] Grinberg, M. (2018). *Flask web development: developing web applications with python*. " O'Reilly Media, Inc.".

[Guo et al., 2019] Guo, W., Gao, H., Shi, J., Long, B., Zhang, L., Chen, B.-C., and Agarwal, D. (2019). Deep natural language processing for search and recommender systems. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3199–3200.

[Harris et al., 2020] Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825):357–362.

[Heijke et al., 2003] Heijke, H., Meng, C., and Ris, C. (2003). Fitting to the job: the role of generic and vocational competencies in adjustment and performance. *Labour economics*, 10(2):215–229.

[Hmoud et al., 2019] Hmoud, B., Laszlo, V., et al. (2019). Will artificial intelligence take over human-resources recruitment and selection? *Network Intelligence Studies*, 7(13):21–30.

[Ho et al., 2019] Ho, A., Nguyen, A., Pafford, J. L., and Slater, R. (2019). A data science approach to defining a data scientist. *SMU Data Science Review*, 2(3):4.

[Hong et al., 2013] Hong, W., Zheng, S., Wang, H., and Shi, J. (2013). A job recommender system based on user clustering. *J. Comput.*, 8(8):1960–1967.

[Inc., 2015] Inc., P. T. (2015). Collaborative data science.

[Indeed, 2021] Indeed (2021). About indeed.

[Indyk and Motwani, 1998] Indyk, P. and Motwani, R. (1998). Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pages 604–613.

[Jiechieu and Tsopze, 2021] Jiechieu, K. F. F. and Tsopze, N. (2021). Skills prediction based on multi-label resume classification using cnn with model predictions explanation. *Neural Computing and Applications*, 33(10):5069–5087.

[JobsPikr, 2019] JobsPikr (2019). 10000 data scientist job postings from the usa.

[Johansson and Herranen, 2019] Johansson, J. and Herranen, S. (2019). The application of artificial intelligence (ai) in human resource management: Current state of ai and its impact on the traditional recruitment process.

[Keuren et al., 2020] Keuren, P., Blatt, D., Ponsen, M., and van den Berg, B. (2020). Wordgraph2vec.

[Kujala et al., 2002] Kujala, S. et al. (2002). *User studies: A practical approach to user involvement for gathering user needs and requirements*. Helsinki University of Technology.

[Lavi et al., 2021] Lavi, D., Medentsiy, V., and Graus, D. (2021). consultantbert: Fine-tuned siamese sentence-bert for matching jobs and job seekers. *arXiv preprint arXiv:2109.06501*.

[Le and Mikolov, 2014] Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR.

[LeVeque, 2013] LeVeque, R. J. (2013). Top ten reasons to not share your code (and why you should anyway). *Siam News*, 46(3).

[Liu et al., 2019] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

[Lletı et al., 2004] Lletı, R., Ortiz, M. C., Sarabia, L. A., and Sánchez, M. S. (2004). Selecting variables for k-means cluster analysis by using a genetic algorithm that optimises the silhouettes. *Analytica Chimica Acta*, 515(1):87–100.

[Luckin and Cukurova, 2019] Luckin, R. and Cukurova, M. (2019). Designing educational technologies in the age of ai: A learning sciences-driven approach. *British Journal of Educational Technology*, 50(6):2824–2838.

[Lund and Gram, 1998] Lund, E. and Gram, I. T. (1998). Response rate according to title and length of questionnaire. *Scandinavian journal of social medicine*, 26(2):154–160.

[MacQueen et al., 1967] MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

[Mäkinen et al., 2021] Mäkinen, S., Skogström, H., Laaksonen, E., and Mikkonen, T. (2021). Who needs mlops: What data scientists seek to accomplish and how can mlops help? *arXiv preprint arXiv:2103.08942*.

[Malinowski et al., 2006] Malinowski, J., Keim, T., Wendt, O., and Weitzel, T. (2006). Matching people and jobs: A bilateral recommendation approach. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*, volume 6, pages 137c–137c. IEEE.

[McInnes et al., 2020] McInnes, L., Healy, J., and Melville, J. (2020). Umap: uniform manifold approximation and projection for dimension reduction.

[Mikolov et al., 2013] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[Miller and Hughes, 2017] Miller, S. and Hughes, D. (2017). The quant crunch: How the demand for data science skills is disrupting the job market. *Burning Glass Technologies*.

[Naili et al., 2017] Naili, M., Chaibi, A. H., and Ghezala, H. H. B. (2017). Comparative study of word embedding methods in topic segmentation. *Procedia computer science*, 112:340–349.

[Nigam et al., 2020] Nigam, A., Tyagi, S., Tyagi, K., and Saxena, A. (2020). Skillbert:"skilling" the bert to classify skills!

[pandas development team, 2020] pandas development team, T. (2020). pandas-dev/pandas: Pandas.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[Peng, 2015] Peng, R. (2015). The reproducibility crisis in science: A statistical counterattack. *Significance*, 12(3):30–32.

[Pennington et al., 2014] Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

[Rani[1] and Rohil, 2013] Rani[1], Y. and Rohil, H. (2013). A study of hierarchical clustering algorithm. *ter S & on Te SIT*, 2:113.

[Reddit, 2021] Reddit (2021). A place for data science practitioners and professionals to discuss and debate data science career questions. `https://www.reddit.com/r/datascience/`. Accessed: 2021-04-25.

[Reimers and Gurevych, 2020] Reimers, N. and Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. *arXiv preprint arXiv:2004.09813*.

[Richardson, 2007] Richardson, L. (2007). Beautiful soup documentation. *April*.

[Rivera et al., 2018] Rivera, A. C., Tapia-Leon, M., and Lujan-Mora, S. (2018). Recommendation systems in education: A systematic mapping study. In *International Conference on Information Technology & Systems*, pages 937–947. Springer.

[Saito and Watanobe, 2020] Saito, T. and Watanobe, Y. (2020). Learning path recommendation system for programming education based on neural networks. *International Journal of Distance Education Technologies (IJDET)*, 18(1):36–64.

[Siting et al., 2012] Siting, Z., Wenxing, H., Ning, Z., and Fan, Y. (2012). Job recommender systems: a survey. In *2012 7th International Conference on Computer Science & Education (ICCSE)*, pages 920–924. IEEE.

[Sobel, 2020] Sobel, B. (2020). Hiq v. linkedin, clearview ai, and a new common law of web scraping. *SSRN Electronic Journal.*

[Tan et al., 2002] Tan, C.-M., Wang, Y.-F., and Lee, C.-D. (2002). The use of bigrams to enhance text categorization. *Information processing & management*, 38(4):529–546.

[Taylor, 2016] Taylor, D. (2016). Battle of the data science venn diagrams. *KDNuggets News.*

[Tiwari et al., 2021] Tiwari, P., Pandey, R., Garg, V., and Singhal, A. (2021). Application of artificial intelligence in human resource management practices. In *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pages 159–163. IEEE.

[Van Der Maaten et al., 2009] Van Der Maaten, L., Postma, E., Van den Herik, J., et al. (2009). Dimensionality reduction: a comparative. *J Mach Learn Res*, 10(66-71):13.

[Van Rossum and Drake, 2009] Van Rossum, G. and Drake, F. L. (2009). *Python 3 Reference Manual.* CreateSpace, Scotts Valley, CA.

[Verma et al., 2012] Verma, M., Srivastava, M., Chack, N., Diswar, A. K., and Gupta, N. (2012). A comparative study of various clustering algorithms in data mining. *International Journal of Engineering Research and Applications (IJERA)*, 2(3):1379–1384.

[Von Luxburg et al., 2012] Von Luxburg, U., Williamson, R. C., and Guyon, I. (2012). Clustering: Science or art? In *Proceedings of ICML workshop on unsupervised and transfer learning*, pages 65–79. JMLR Workshop and Conference Proceedings.

[Wilkinson et al., 2019] Wilkinson, W., Podhorska, I., and Siekelova, A. (2019). Does the growth of artificial intelligence and automation shape talent attraction and retention? *Psychosociological Issues in Human Resource Management*, 7(1):30–35.

[Yorke and Knight, 2006] Yorke, M. and Knight, P. T. (2006). *Embedding employability into the curriculum*, volume 3. Higher Education Academy York.

[Zhang et al., 2010] Zhang, Y., Jin, R., and Zhou, Z.-H. (2010). Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, 1(1-4):43–52.

[Zhang et al., 2018] Zhang, Y., Lu, J., Liu, F., Liu, Q., Porter, A., Chen, H., and Zhang, G. (2018). Does deep learning help topic extraction? a kernel k-means clustering method with word embedding. *Journal of Informetrics*, 12(4):1099–1117.

Skill Scanner: CV Review for JOB SEEKERS

Skill Scanner used AI to compare your skills:

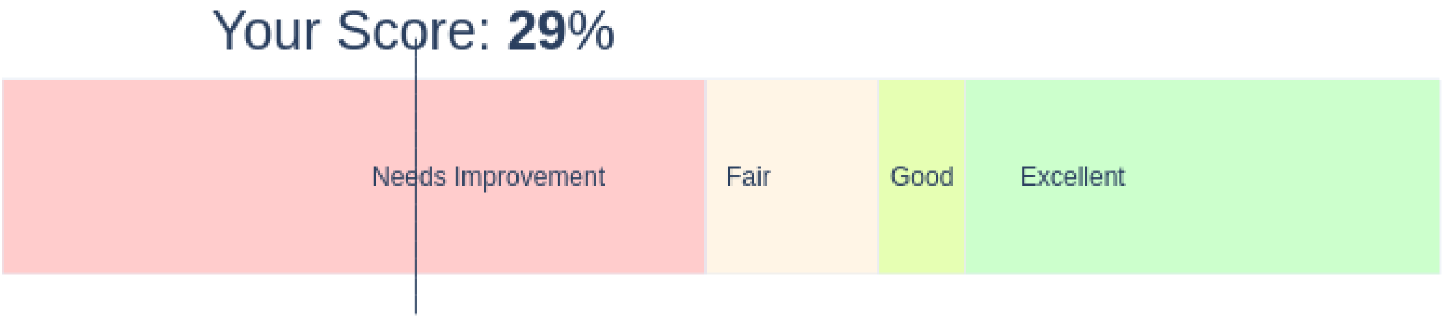
1. Comparison to Competition: your skills compared to representative CV's.
2. Fit to Demand: Insight in your skill gaps from comparing your CV to employer demands.
3. Find and Select: Recommendations for Education to fill your skill gaps.

**For an in depth explanation of our technique please refer to the last page.

1. Comparison to Competition

See how you rank up against other JOB SEEKERS' CV's.

Your total CV Coverage Score:



What do the scores mean?

Outperformed by over 50% of JOB SEEKERS

Outperforms 50% of JOB SEEKERS

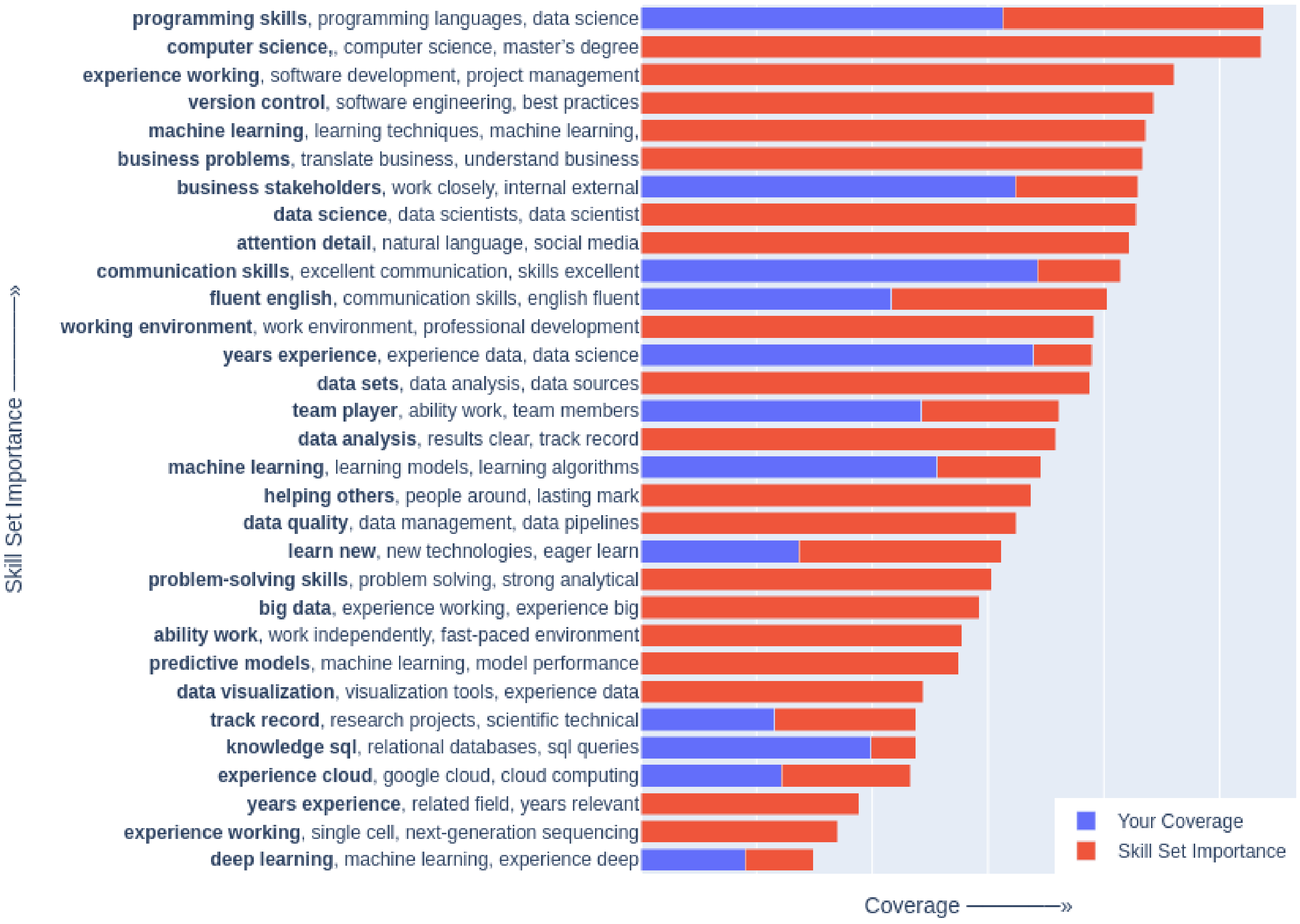
Outperforms 75% of JOB SEEKERS

Outperforms 90% of JOB SEEKERS

*** Note: In this example you can only input 5 skills, with more input skills the coverage score will increase.*

2. Fit to Demand

Gain insight in your skill gaps: See how well your CV covers the most important skill sets demanded in Job Postings.



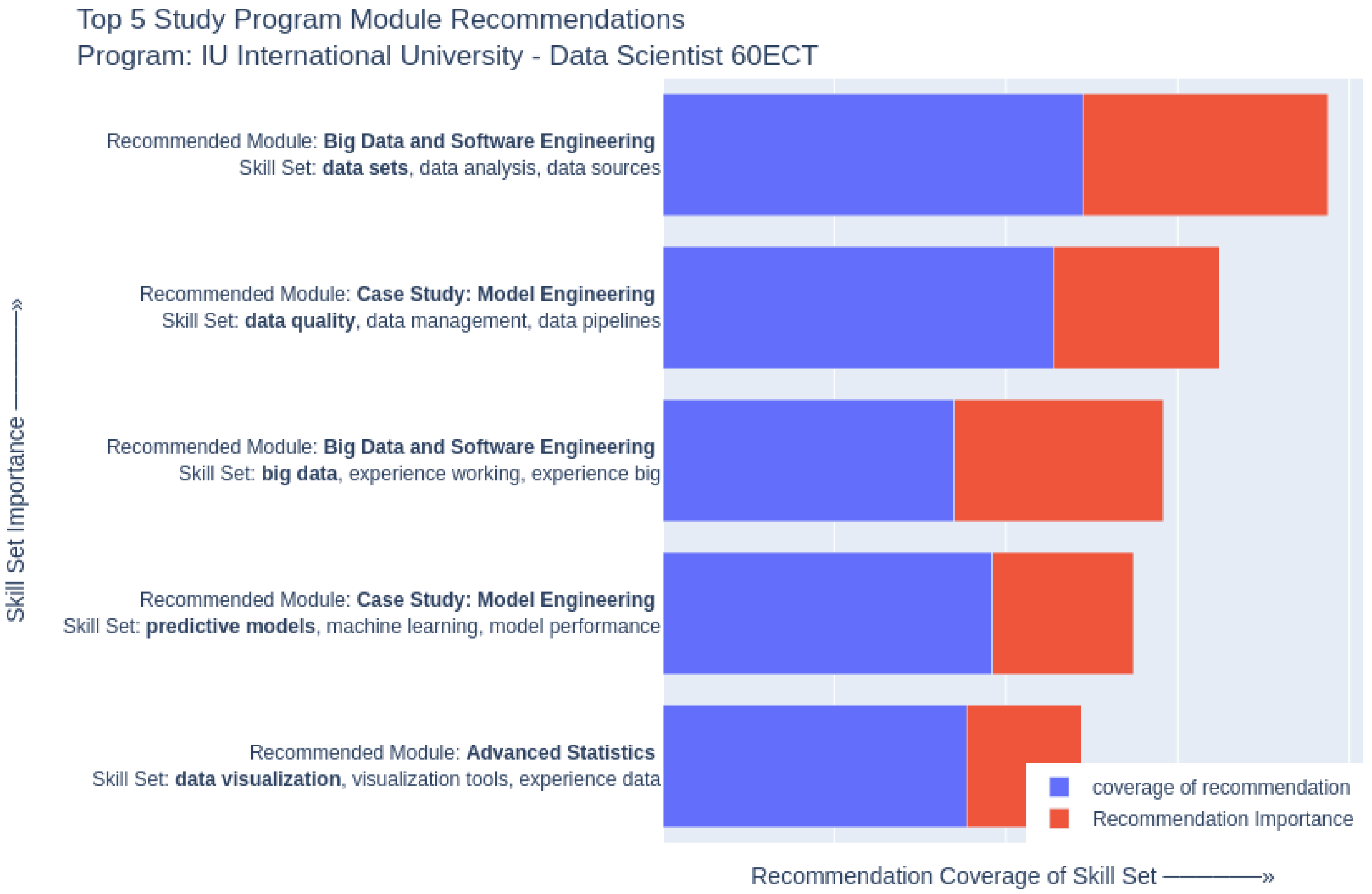
*** Note: A coverage of 100% is impossible to attain, a coverage of over 70% can be considered excellent.*

3. Find and Select

Choose the right study program for YOU

Are you considering formal education to upskill yourself? Skill Scanner helps you find the right program:

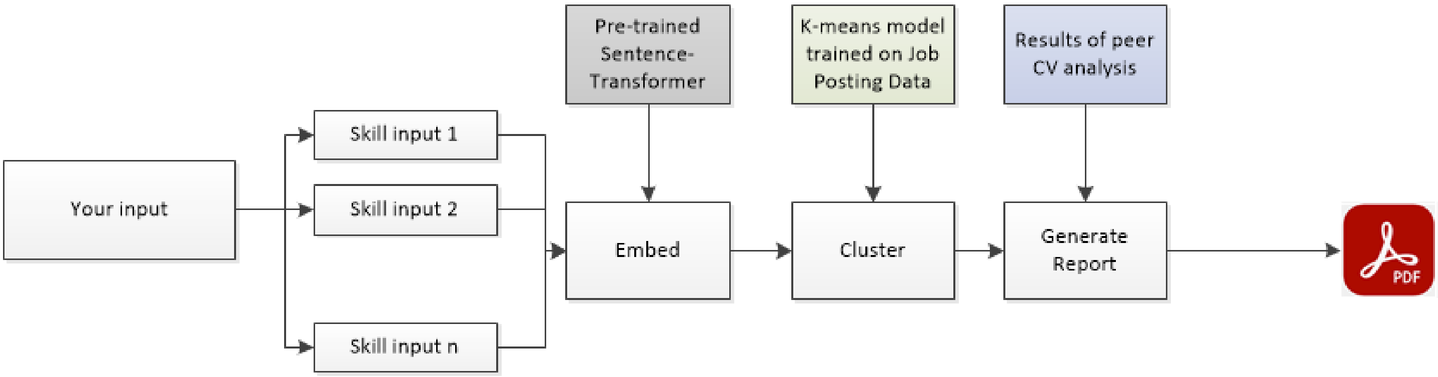
- We exposed skill gaps in your CV.
- We analyzed which training courses complement your CV.
- You find the top 5 recommendations in the figure below.



*** Note: In the future more learning content will be analyzed.*

Thank you for reading, please remember to fill out our questionnaire at https://forms.gle/ct4DSno6UxN4qofu8

Appendix A: Methodology Explanation


Pre-trained Sentence-Transformer
K-means model trained on Job Posting Data
Results of peer CV analysis
Your input
Skill input 1
Skill input 2
Skill input n
Embed
Cluster
Generate Report
PDF

Functionality sketch of Skill Scanner backend

Skill Scanner processes your input skills in three steps:
1. Embedding: We use Sentence-Transformers, a state-of-the-art framework for sentence embeddings. In simple terms, a sentence embedding algorithm turns a sentence into a series of numbers from which a computer can infer differences and similarities.

2. K-means modelling: The sentence embeddings from step 1 enable us to compute the similarity of various embeddings. A K-means model was used to cluster skills in groups of similar meaning.

2.1 Training of K-means model: The training data is a dataset of 21.500 job requirements extracted from Data Scientist vacancy retrieved from various sources.

2.2 Evaluation of K-means model: To evaluate how the model generalizes to different data sources we inferred clusters from skillsets found in a sample of Data Scientist CV's. The model was able to infer the correct cluster with an accuracy of 82%.

2.3 Use of the K-means model by Skill Scanner: Skill Scanner uses the K-means model to infer clusters from you input data. It uses cosine similarity to compute the distance from the cluster centroid. This metric is reported as similarity score.

3. Peer CV Analysis: We used Skill Scanner to analyze a dataset of 65 Data Scientist CV's retrieved from Kaggle.com. The results of this analysis allows us to compare your scores to these of your peers (other Data Scientists).