

# Combining Embeddings and Domain Knowledge for Job Posting Duplicate Detection

Matthias Engelbach<sup>1,\*</sup>, Dennis Klau<sup>2</sup>, Maximilien Kintz<sup>1</sup>, Alexander Ulrich<sup>3</sup>

<sup>1</sup> Institute for Industrial Engineering IAO, Fraunhofer,  
Germany

<sup>2</sup> University of Stuttgart,  
Institute of Human Factors and Technology Management IAT, Stuttgart,  
Germany

<sup>3</sup> Contractor Consulting GmbH,  
Germany

{matthias.engelbach, maximilien.kitz}@iao.fraunhofer.de,  
dennis.klau@iat.uni-stuttgart.de, alexander.ulrich@contractor.de

**Abstract.** Job descriptions are posted on many online channels, including company websites, job boards or social media platforms. These descriptions are usually published with varying text for the same job, due to the requirements of each platform or to target different audiences. However, for the purpose of automated recruitment and assistance of people working with these texts, it is helpful to aggregate job postings across platforms and thus detect duplicate descriptions that refer to the same job. In this work, we propose an approach for detecting duplicates in job descriptions. We show that combining overlap-based character similarity with text embedding and keyword matching methods lead to convincing results. In particular, we show that although no approach individually achieves satisfying performance, a combination of string comparison, deep textual embeddings, and the use of curated weighted lookup lists for specific skills leads to a significant boost in overall performance. A tool based on our approach is being used in production and feedback from real-life use confirms our evaluation.

**Keywords.** Job posting analysis, similarity embeddings, domain knowledge, duplicate detection, deployed application.

## 1 Introduction

With increasing digitization of business processes, the possibility to assist people working in the area of human resources by leveraging means of automated recruiting has become an important topic [4, 14]. Technologies like artificial intelligence (AI) and robot process automation (RPA) can be applied to a wide range of tasks in a typical recruiting process, such as automated screening of job portals (e.g., LinkedIn) for retrieving job offerings published by hiring companies or personal job profiles and CVs [13].

However, even with advanced methods and tools for collecting such data, further processing can be challenging. Especially when merging data stemming from multiple sources, curating and cleaning the vast amount of records captured regularly is crucial [5, p. 6]. For instance, a database might contain current job postings targeting recruitment of freelancers for IT projects that have been published on several job platforms in parallel and which in fact relate to one single job reference. As of now, while many other steps in

the recruitment process can already be automated quite successfully, the detection of such duplicates within unstructured textual data is not an easy task to solve [20]. This is partially due to the fact that platform and company specific constraints often result in similar, but not identical postings for both: different job advertisements being part of the same project, as well as job postings coming from the same company, but relating to different projects [2].

For those reasons simple automated duplicate checks usually lead to a high rate of false positives, which is why this task is often still solved in the form of time consuming manual work. To address this problem and to provide a useful programmatic solution, we investigated different methods of similarity detection tailored to the specific goal of finding duplicates among the textual contents of job postings.

Our key findings are that several approaches deliver good results, but only a combination of text similarity matching with domain knowledge in the form of a weighted skills averages and text embedding gives a very high score with few false positives. An application based on our duplicate detection approach was developed in a economically viable way and has been successfully deployed as a live system for the past months.

## 2 Industrial Use Case and Requirements

Recruiters deal with the matching of candidates with job positions, for example finding a suitable IT freelancers for specific software development projects. One task the recruiters need to perform consists in the curation of a job listings database. Companies looking for IT freelancers often publish open positions on many different web sites, in slightly different variations.

These open positions need to be aggregated and duplicates need to be identified. This time-consuming process occurs mostly manually or with some degree of automation that however is error prone and requires manual oversight. To give an idea of the difficulty of the task, some example borderline cases can be named:

- the same job posting may be published on two different sites with slightly different wording in the title and text description,
- the same company may be looking for two people for the same Java development project, one for front-end and one for back-end, resulting in very similar postings for different jobs or,
- the same employer may be looking for two Java developers in a project, one working on site in the company offices, one being a remote working position.

These examples show that detecting duplicate job listings is hard to achieve in a reliable way by simply matching basic data such as job title description and company name, or only analyzing the required skills. A sophisticated combination of several different criteria, however, can greatly increase the detection performance.

As the project was conducted for a real-world use-case in cooperation with Contractor Consulting GmbH<sup>1</sup>, the goal was to have a tool ready for production use. Because of this, some specific requirements applied that may would not have been relevant in the setting of a larger research project:

- Both, the development and operation costs had to be limited in order to get a viable business case. The development and usage of the resulting tool must be cheaper than the labor cost of manually detecting duplicates, and the additional costs generated by false detection (false positives/negatives). The tool will be used in a semi-automated fashion, where listings marked as duplicate, i.e. true and false positives, are reviewed by the recruiting experts; while non-duplicates are processed automatically and no further review or action is triggered. Therefore, in the context of misclassifications, false positives pose less of an issue to the users than false negatives due to the human review and feedback loop. The potential savings generated by the new tool do not justify the development and usage of large-scale solutions. Development effort was limited to about 15 to 20 person days.

<sup>1</sup>[www.contractor.de](http://www.contractor.de)

- Additionally, the software shall have no licensing and hardware renting costs. This means that using large cloud AI platforms or hardware-intensive large language models was not an option.
- Whenever possible, easy to understand and easy to maintain solutions should be preferred, so as to ensure that the tool can be maintained by the end users.

These requirements constrained the scope of considered approaches and tools used during the project, also excluding the fine-tuning of big language models with curated data sets.

### 3 Related Work

Usage of AI solutions in recruitment processes is common and well documented, though typically the task addressed is the analysis and review of applications and resumes [3].

Some research has already addressed job descriptions. In [27], the authors show how removing stop words and extracting relevant features from online job description postings help in the detection of fake postings.

Works like [2] address the specific task of duplicate detection for online recruitment using standard de-duplication techniques like Jaccard Similarity and n-grams with limited success. Furthermore, [30] investigated different 24 methods for job duplicate detection, including usage of different tokenizers.

Another common use case from a different, yet still technical domain is the detection of duplicates among free text bug reports, where in some cases similar techniques have been applied [25] [26]. However, extensive usage of domain knowledge is not applied there.

Text comparison is a common task for which many approaches exist. However, classic algorithms such as Levenshtein distance [15], TF-IDF [24] or longest common subsequence (LCS) [16] are not suitable for our approach, as the job listings contain many similar-looking text blocks on a word and character basis, but are often very different in terms of meaning.

[10] use Word2Vec [17] embeddings to detect short text duplicates. The usage of embeddings is clearly relevant to our work, job descriptions tend to be long documents and thus require a different approach than the one proposed by the authors.

[9] provide a review of plagiarism detection methods, an area related but not identical to our problem domain. The findings confirm that using new machine learning techniques and a combination of approaches is most promising.

[28] give an overview of text similarity measurement techniques. One main finding of the study that is relevant to the present work is the confirmation that text matching is application and domain-specific, i.e. dependent on the type of texts and on the task to be accomplished after the similarity has been computed. In order to create performant solutions, different algorithms need to be considered.

Overall, our overview of related work shows that though many approaches exist for computing text document similarity, a selection of the most appropriate one is always task and domain specific and thus needs to be reviewed for each new project. The particularities of job descriptions texts and the project requirements mentioned in the previous section lead us to conduct the experiments described in the following chapters.

### 4 Data Resources

During the project phase we had access to a database with about 45.000 collected project tenders and according textual descriptions. The typical content of a job description is well established and has already been studied. [18] mention the main elements of such a text: a job title, required skills, preferred skills and responsibilities. In our specific case and dataset, we identified the following features as most relevant:

- **Title:** The headline on top of the project proposition.
- **Description:** The text body describing the project and job requirements.

Job Description Text 1	Job Description Text 2
<p>Aufgaben Aktuell sind wir für einen unserer Kunden auf der Suche nach einem IT Lead for SAP S/4 HANA template extension project Zu Ihren Aufgaben zählt Ensuring the integrity and future viability of the global template with a focus on the SAP/4HANA for Retail solution Analysing and assessing business requirements in cooperation with a team of Solution Architects/ System Integrator Managing external partners during the Tendering Selection ampContracting phase Provide Input and estimation model for Resource and Budget forecast Projekt-Nr. [redacted] Projekttitel IT Lead for SAP S/4 HANA template extension project m/w/dStellentyp freiberuflichEinsatzort D4Starttermin asapDauer 12 Monate Anforderungen Must-have Degree or similar qualification apprenticeship with a focus on IT/business administration Project experience in implementing or managing SAP S/4HANA Retail template projects Process and SAP knowledge in a supply chain context especially buying and logistics in the retail sector Extensive experience in realizing technical solutions development methods and different test and rollout approaches Zusätzliche Informationen Konnten wir Ihr Interesse wecken Dann freuen wir uns auf die Zusendung Ihres aussagekräftigen Expertenprofils unter Angabe Ihrer Stundensatzvorstellung [redacted] zählt zu den führenden Technologiedienstleistern in den Bereichen Softwareentwicklung IT [redacted]</p>	<p>Aufgaben Aktuell sind wir für einen unserer Kunden auf der Suche nach einem Project Manager for SAP S/4 HANA template extension project Zu Ihren Aufgaben zählt Implementing project management processes such as project planning and financial management Supporting and advising the project lead an project management methods processes and tools Project Status Reporting an Management Level reale update and align budget and resource forecasts Cooperating with our international Offices and partners Assuming the overall I coordination with functional and expert teams Projekt-Nr. [redacted] Projekttitel Project Management for SAP S/4 HANA template extension project m/w/dStellentyp freiberuflichEinsatzort D4Starttermin asapDauer 12 Monate Anforderungen Must-have Experience in international project management of SAP transformation projects Ideally excellent knowledge of MS Office SharePoint and LIRA Excellent numerical skills and analytical thinking Diplomatic skills team spirit flexibility and assertiveness Zusätzliche Informationen Konnten wir Ihr Interesse wecken Dann freuen wir uns auf die Zusendung Ihres aussagekräftigen Expertenprofils unter Angabe Ihrer Stundensatzvorstellung [redacted] zählt zu den führenden Technologiedienstleistern in den Bereichen Softwareentwicklung IT [redacted]</p>
Skill Text 1	Skill Text 2
<p>lead sap s/4 hana template project integrity template focus sap retail solution business requirements cooperation team solution system integrator tendering estimation model budget forecast lead sap s/4 hana template project qualification focus business administration project experience sap s/4hana retail template projects process sap knowledge supply chain buying logistics retail experience technical solutions development test rollout engineering</p>	<p>project manager sap hana template project project management processes project planning financial management project lead project management processes tools project reporting management update budget forecasts international coordination expert teams project management sap hana template project experience international project management sap projects knowledge ms office sharepoint skills skills team spirit engineering</p>

**Fig. 1.** Example of job offering comparison for two different positions that are part of the the same project. While the original description text has greater parts of overlap in boiler plate text sections, the extracted pure skill texts differ significantly. Yellow markings indicate overlapping text parts between source and target texts. The censored parts mark confidential information

Additionally, each project in the data set is identified by a unique project ID. The *description* is the textual main body of the job posting and typically consists of the following elements:

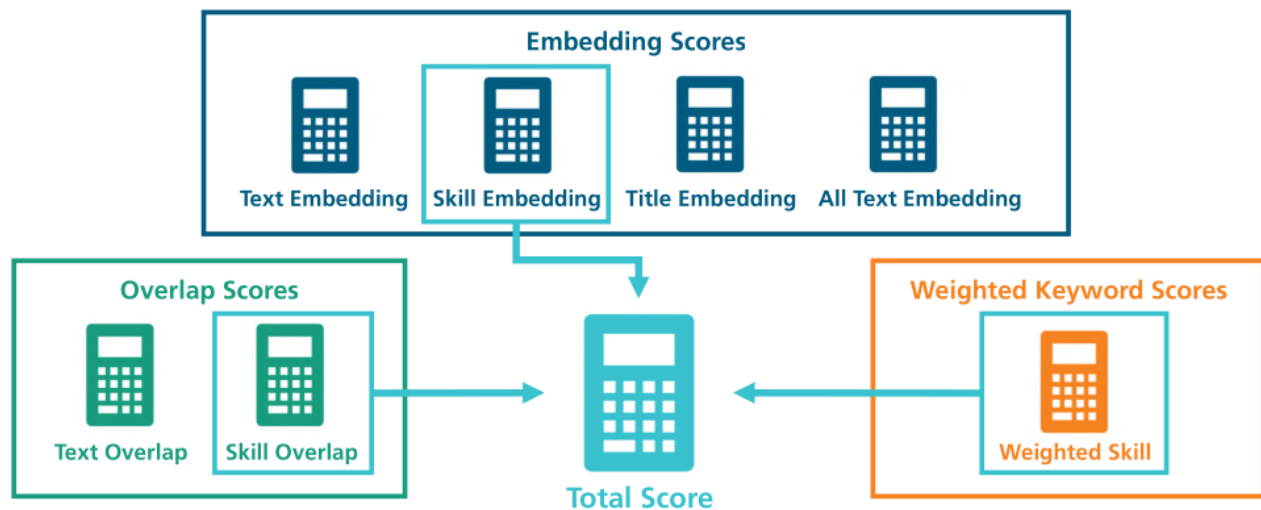
- a short *introduction* text usually describing the company that is looking for a freelancer,
- the *tasks to be accomplished* – often in form of a bullet-point list,
- *requirements to be met* by the potential candidate, for example years of experience in a specific area, required hard and soft skills, etc.,
- *general working information* such as start and end date of the project, remote or on-site position, or amount of hours of work expected,
- a short *ending* text describing the application process.

The exact elements and their order may vary from document to document. Additionally, many postings contain additional "glue" or "boilerplate" text added by the company or recruiting agency, which does not carry information for our task, but can lead to false positive detections.

Besides the textual information that we get for each job offer, the data set also contains some meta information for each sample like date of publishing, which we use to limit matching candidates to a realistic time range in which we expect duplicates of the same job to appear (e.g., a few weeks).

Other features like working place, duration (for freelancers that work on specific projects this is usually already fixed in the job posting) and starting date are also stored for every job posting. Although these structured features could also be indicators for duplicates, during our pretesting we found them to be misleading for many cases as well, since different projects and jobs might share the same place or dates and some features are missing for some candidates in the database.

Even though such information might be useful for the validation of detected duplicates, in our work we focus on comparing the larger textual features mentioned before such as job description and title. In addition to the job posting data, we had access to an (automatically created) key word list containing the 25.000 most frequent skills (occurring at least seven times in resumes).



**Fig. 2.** Overview of the different scoring approaches in their respective matching classes and the final selected component of each class for the total score. The components were selected by their performance on the validation set

Accompanied by the skill list we also introduced a *skill black list* during the project phase with the aim to exclude misleading or wrong entries in the skill list. A small test set was manually annotated with 50 *duplicate-pairs* identified by Contractor. This test set exhibits a high quality and was used only for final testing of the deployed pipeline.

After the implementation of the first version of the detection pipeline, we created a validation set by annotating a set of 176 job posting-pairs with corresponding predictions by the pipeline. After curation, we ended up with 74 duplicates and 73 non-duplicates.

We refer to those 73 non-duplicates also as *fake-duplicates*, since they were labeled as duplicates by our algorithm and therefore exhibit at least some degree of similarity to the matching source.

29 of the predictions could not be assigned confidently to one of the classes (duplicate, non-duplicate) even after review from the domain experts and were removed from the curated dataset. This validation set was used to fine-tune the parameters and thresholds during development of the matching algorithms.

## 5 Method and Experimental Setting

### 5.1 Data Preprocessing

Before comparing job offerings with each other, the whole dataset (and especially the textual contents) are preprocessed and cleaned: we remove special characters such as line breaks or enumeration characters to receive a normalized version of all texts – mainly containing letters, white spaces and numbers. We make an exception for special characters like '+' or slashes that can appear in technical skills (e.g., C++, S/4 Hana), which we keep in the text corpus.

Besides, similar to the approach proposed by [12], we extract and lowercase all skill terms from the original description texts based on the skill list and skill black list mentioned in the previous section together with their weights (see Section "Weighted Keyword Matching"). This preprocessing is done for all further matching methods and score computations.

### 5.2 Evaluation Metrics

For measuring the quality of our matching algorithms, we resort to standard metrics like Precision, Recall and F1-Score, which have proven to be reliable indicators for duplicate detection

tasks [1]. We compute them over all matched pairs of project publication records in our data sets. In detail, we handle a comparison result as true positive (TP) if an actual duplicate has been detected successfully, and as false negative (FN) if this is not the case. Similarly, non-duplicate pairs that have been assigned with a duplicate label are treated as false positives (FP), while true negatives (TN) are all pairs of correctly labeled non-duplicates.

As proposed by previous works [7], we use score dependent threshold values (TH) for the decision of label assignment. Thus, if a matching pair sample gets assigned a score above the threshold by an algorithm, it is labeled as a predicted duplicate, otherwise it is assigned the non-duplicate prediction label.

We use the validation set containing both duplicates and fake-duplicate job pairs introduced in the previous section as a development set for score and threshold optimization and evaluate our final pipeline on the test set containing only known duplicate pairs to ensure that our approach is valid for unseen data.

### 5.3 Overlap-based Similarity Matching

We use a simple *Text Overlap Score* computation, that computes the degree of matching text blocks between two job publications and serves as a baseline for all further experiments. In detail, the Python package *levenshtein*<sup>2</sup> is used to get all larger text blocks (with minimum length of 15 characters) within the source text that also appear in the target candidate of the current matching pair.

We define  $S$  as the source text and  $T$  to be the target matching text. For each overlapping block  $i$ , we define  $b_i$  as its length. If there are a total of  $n$  overlapping blocks, the total length of overlapping text is:

$$B = \sum_{i=1}^n b_i. \quad (1)$$

Then, the  $TOS$  is calculated as follows:

$$TOS = \frac{B}{\text{len}(S)}. \quad (2)$$

<sup>2</sup>[maxbachmann.github.io/Levenshtein/levenshtein.html](https://maxbachmann.github.io/Levenshtein/levenshtein.html)

This means, we sum up the length of the overlapping string blocks (number of overlapping characters) and normalize the result by dividing through the total length of the source text to get a percentage value that indicates the degree of overlap.

During our first pretest experiments using the overlap score, we found that many of the job offering texts consist of larger blocks of boiler plate text (such as introduction of the company or common contact information) that are shared among different job offerings published by the same company or agency.

This results in high  $TOS$  values for different job postings. On the other hand, we identified other passages of the description text as most relevant for distinguishing between job offers, specifically all sections containing project-related skill requirements.

In Figure 1 we illustrate this finding with an example: The start and end of the matching source text overlap in most parts with the target, while the skill list indicates that in this case indeed different job offerings are compared.

For this reason, besides computing a simple overlap score over the whole job description text, we additionally compute a *Skill Overlap Score* ( $SOS$ ) that only considers matching blocks appearing in the concatenation of all listed skill words and terms detected in the whole description text (see bottom of Figure 1). These pure skill texts, representing the most specific parts of a job offering, are extracted during the preprocessing phase by using the skill list.

Note, that for each pair matched with the overlap scores we compute a forward and a backward score, since the choice of source and target texts influences the length of the resulting matching blocks. Therefore, the final  $TOS$  is calculated by taking the mean of the forward and backward pass:

$$TOS_{\text{final}} = \frac{TOS_{\text{forward}} + TOS_{\text{backward}}}{2}. \quad (3)$$

The final Skill Overlap Score  $SOS$  is computed analogously.

**Table 1.** Experimental results for different matching score computations listed as Precision (P), Recall (R) and F1-Score (F1) of correctly recognized project duplicates on the validation set. TH indicates the threshold chosen for the specific score, and \* the scores used to compute the Total Score (*TS*)

Matching Score	TH	P	R	F1
Text Overlap ( <i>TOS</i> )	0.40	0.48	0.66	0.55
Skill Overlap ( <i>SOS</i> ) *	0.40	0.89	0.66	0.76
Text Embedding ( <i>TES</i> )	0.50	0.52	0.81	0.63
Skill Embedding ( <i>SES</i> ) *	0.50	0.66	0.97	0.79
Title Embedding ( <i>TTES</i> )	0.50	0.49	0.74	0.59
All Text Embedding ( <i>AES</i> )	0.50	0.57	0.95	0.71
Weighted Skill ( <i>WSS</i> ) *	0.20	0.84	0.97	0.90
<b>Total Score (<i>TS</i>)</b>	<b>0.35</b>	<b>0.89</b>	<b>0.99</b>	<b>0.94</b>

#### 5.4 Embedding-based Similarity Matching

In addition to the string overlap scores, we also test similarity matching based on text embeddings in order to capture semantic similarities (e.g., different wordings and phrases) not detected by the strict string matching approach. In detail, since our data contained both German and English job descriptions, we used a multilingual text encoding method as proposed by [29] and provided by the pretrained model *distiluse-base-multilingual-cased-v1*<sup>3</sup> from the Python library *sentence-transformers* [21].

Embeddings for all data samples are computed during preprocessing and cached for all further comparison actions. During pairwise matching of job descriptions, we compute the *Embedding Score (ES)* with the commonly used cosine similarity [23] between embedding vectors  $S_e$  of source  $S$  and  $T_e$  of target candidate  $T$  as follows:

$$ES = \text{cosine similarity}(S_e, T_e) = \frac{S_e \cdot T_e}{\|S_e\| \|T_e\|}. \quad (4)$$

We then normalize the results by mapping (eventually appearing) negative values to 0 (indicating no significant similarity between  $S$  and  $T$ ) for ensuring comparability of scores.

<sup>3</sup>[www.sbert.net/docs/pretrained\\_models.html](http://www.sbert.net/docs/pretrained_models.html)

Similar to the computation of the overlap scores, we calculate embedding scores for different parts of the original job description:

- *Text Embedding Score (TES)*: Embedding over the whole job description text.
- *Skill Embedding Score (SES)*: Embedding over the pure skill text (concatenation of all found skill terms).
- *Title Embedding Score (TTES)*: Embedding over the headline of the job description (retrieved from a separate feature in the data set).
- *All Text Embedding Score (AES)*: Embedding over concatenation of all texts (title, description and skill text).

#### 5.5 Weighted Keyword Matching

During the pretesting phase of the project, we found that skill terms mentioned in job descriptions are key features regarding the task of duplicate detection. On the other hand, since the list of relevant skills was created in an automated manner, it also contains wrong or misleading keywords (examples for that are "English", "experience" or "team work") that are not indicating real technical skill requirements for the job.

Furthermore, very common terms like "Java", that are occurring in a lot of job descriptions, are considered as not helpful for finding duplicates. Although a skill black list as introduced in the Data Resources section was used, manual capturing of all misleading terms was not feasible within the scope of the project.

For those reasons and as proposed by other work like [12], we introduce another matching approach: a *Weighted Skill Score (WSS)*. In contrast to the previous, presented scores which use the skill list to different degrees, the *WSS* focuses on the significance of individual skill keywords by taking into account their relevance among the whole corpus of the data set. For that, we calculate weights for every skill term in the skill list during preprocessing by counting their occurrences in all project descriptions of the data set to approximate a value of their significance.

The weights are re-computed on a regular basis, which could also be done in an online manner while the system is active and new project descriptions are added to the data base. This way the system can continuously learn during its operation time. For a skill term  $s_i$  from the skill list, we define  $f(s_i)$  as its frequency in the entire project database. Then, the weight of the skill term is:

$$w(s_i) = \frac{1}{f(s_i)}. \quad (5)$$

During calculation of the weighted score for a matching pair, we take the inverse of every precomputed skill term's frequency because we assign more frequent skill words less relevancy for duplicate detection than specific skills that appear very seldom. For a source text  $S$  and a target candidate  $T$ , the *Weighted Skill Score*  $WSS$  is then calculated as follows:

$$WSS = \frac{\sum_{s_i \in (S \cap T)} w(s_i)}{\sum_{s_i \in S} w(s_i)}. \quad (6)$$

The  $WSS$  consists of the sum of weights of all skill terms found in both job descriptions  $S$  and  $T$  divided by the sum of weights of all skills within the source text of the matching pair. Again, we perform a forward and backward pass by swapping  $S$  and  $T$  and use the mean of both for the final weighted skill scores  $WSS_{\text{final}}$ :

$$WSS_{\text{final}} = \frac{WSS_{\text{forward}} + WSS_{\text{backward}}}{2}. \quad (7)$$

## 5.6 Combined Matching

As a result of the experiments during a first pretesting phase, we deduced a final combined *Total Score* ( $TS$ ) that is calculated by averaging over the three (best performing) individual scores – each coming from one of three categories mentioned before.

Figure 2 visualizes the score influences contributing to our total score: we found that the skill related metrics perform best during our tests, thus the  $TS$  consists of the average of Skill Overlap Score  $SOS$ , Skill Embedding Score  $SES$  and Weighted Skill Score  $WSS$ .

In principal, all employed scoring methods are language and location independent and have been applied to German and English job postings. For better performance of the embedding methods, it is advised to use an appropriate embedding model for the target language or a multilingual one.

## 6 Results and Discussion

### 6.1 Threshold Finding

To identify suitable thresholds for each score, we start visualizing the prediction results for the different methods on the validation set with duplicates and (similar looking) non-duplicates. Figure 3 shows those predictions for every scoring method introduced in the previous section.

Note that, due to the constrained scope of the project, we set the threshold values manually by human supervision to achieve a good prediction performance for each scoring method. However, we use the same values for every method group to preserve comparability. The thresholds are listed in Table 1 together with the according score performance when applying these threshold values for prediction on the validation set.

### 6.2 Experimental Results

Comparing the performance of individual scores, Table 1 and Figure 3 indicate that using the text overlap ( $TOS$ ) and embedding score ( $TES$ ) over the whole description texts is not sufficient for discriminating between real and fake duplicates. This is also due to the already mentioned existence of overlapping boilerplate text: some texts originate from the same company or even project, but indeed advertise different jobs. When concentrating on the skill text containing only relevant technical terms specific for a given job offer by using the  $SOS$  and  $SES$  metrics, performance increases significantly.

In contrast, relying on the individual result of the title score  $TTES$  is not helpful. Although in the beginning we expected the headline of a job offering to be a meaningful indicator for duplicate references, we realized that these texts without any further context are also highly misleading, since a lot of independent companies and projects



use the same wording and technical terms (e.g., "SAP developer wanted") when publishing job postings. Even combining the title and skill text with the original job description text in the all text embedding approach (*AES*) cannot deliver good performance.

Overall, one of our key findings was that – although they achieve the highest prediction scores for real duplicate pairs – all of our embedding approaches in general seem not to be capable of separating fake duplicates from real ones, as the corresponding plots in Figure 3 underline.

On the other hand, only using the weighted skill score (*WSS*) without combining with any other measure already performs well. The fact that we attribute weights to significant skill features allows reliable detection of fake duplicates, since in contrast to the other metrics the influence of overlapping (but maybe misleading) boilerplate text on the prediction is reduced.

Note that a requirement for the productive usage of this score computation is the continuous update of the underlying skill weights to capture the relevance of new terms and wordings that might arise over time.

With combining the most promising individual scores of each group (namely the skill based scores) to form the total score *TS*, we can improve the performance on the validation set to an F1 score of 0.94, meaning that most of the comparison pairs could be classified correctly into duplicates and non-duplicates.

We believe that assembling the our scoring method in this way enables us to compensate for individual score weaknesses and to receive more reliable and stable prediction results in the combined score, as shown before [8]. During the final phase of the project, we applied this optimized setting to the test set consisting of 50 samples which are exclusively real duplicate pairs.

Since there are no non-duplicates in this dataset, the final test before productive deployment ensures that the constructed total score can identify the actual duplicates well by assigning them a relatively high score. Although there was a clearly visible scatter among predictions with a *TS* ranging from 0.24 to 0.93, we could identify some indicators of similarity for every test sample,

indicating the potential for further fine-tuning of the *TS*. For a threshold of 0.2, all of the real duplicates would have been detected. After feedback from the project partner, however, thresholds were not set to the optimal operating point for the (rather small) test set, but to a more robust value (see next section) for the final productive deployment. This threshold choice implies more importance to specificity (i.e. low false alarm rate) at the cost of sensitivity.

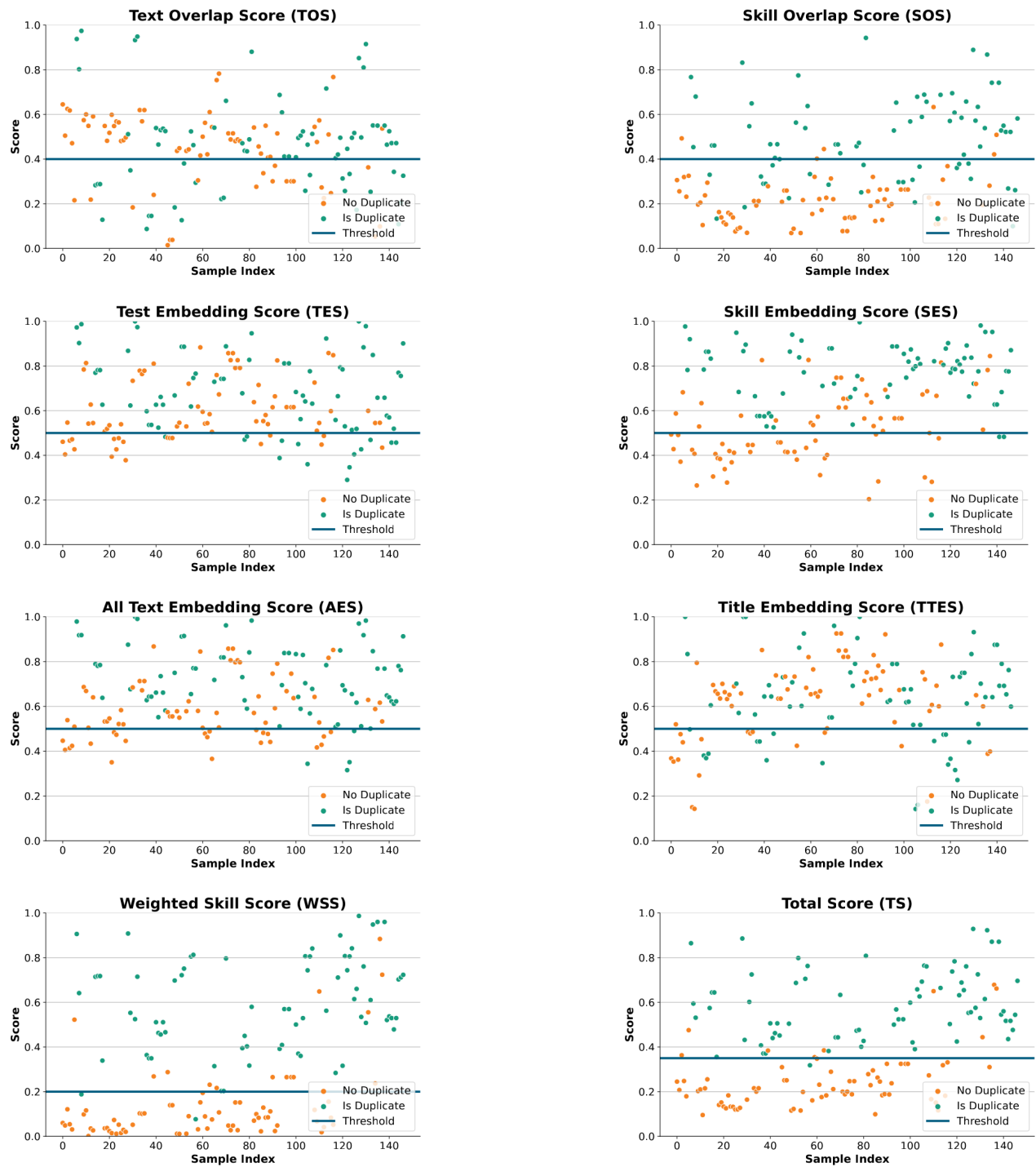
## 7 Productive Deployment

After the experimental evaluation, the duplicate detection framework is deployed in production with periodic checks for recently published and collected job descriptions. In terms of scalability, once the initial vector database for the document embeddings is generated, the computational cost and time scales linearly in the number of processed documents.

Since only a fixed time frame is compared to a new job posting, the number of comparisons can in fact vary but does not scale with the size of the dataset. For the deployed system, the thresholds for the score computations were adapted and an additional criteria introduced: the *TS* threshold is set to 0.6 and each individual scoring method must achieve a minimum value of 0.1 in order to classify a pair as duplicates.

With this setting, 58.860 published job descriptions were analyzed and matched during a first productive phase of six months, of which 33.865 were found to be no duplicates and 24.995 to be duplicates of other projects. In detail, for every sample predicted as a duplicate we identified about 2.5 duplicate job offerings on average.

This implies that the 24.995 duplicate predictions contain about 10.000 unique job postings, leading to a total sum of 43.705 postings identified as unique within this period of time. Overall, the positive feedback given by the end users of the tool after several months of productive usage confirm that our solution is working as intended and validates our approach.



**Fig. 3.** Score predictions for the different methods on the test set with 74 duplicates and 73 non-duplicates. The thresholds are set manually for each score to gain a good distinction between positive and negative duplicate predictions

## 8 Limitations

As stated by [6], choosing thresholds on a small test set of samples might not scale appropriately when the data set gets larger. Therefore, future experiments on larger (labeled) test sets are necessary for a more generalizable threshold estimation. This was however not possible in the scope of this work, due to the nature of requiring high-quality labeled data, which had to be collected manually by recruiting experts during the data collection phase of the project.

An additional source of improvement is a more sophisticated determination of the weights assigned to the individual metrics that form the Total Score ( $TS$ ). The assumption, that individual scoring approaches carry more information than others to solve a task in a specific domain, is reasonable and was also shown in [8].

## 9 Conclusion and Future Work

Our work showed that for duplicate detection in real world scenarios, like detecting clones of job descriptions, the integration of domain specific expert knowledge is crucial. While applying standard matching algorithms like string similarity and embedding comparison only leads to limited performance, we pointed out that combining these approaches with context specific additions – in our case the emphasis of significant skill terms – achieves satisfactory results (F1-score of 0.94) regarding capabilities for productive deployment.

For future work, we plan to do a larger evaluation on bigger (human-curated) test sets to identify possible weak points of our approach that could not be addressed within the limited scope of this first project. Furthermore, as large language models (LLMs) become more advanced even in not explicitly shown domains, they can be used as a baseline classifier in zero- or many-shot evaluations. This allows for a more comprehensive evaluation, with the LLM serving as a proxy for a human evaluator. Additionally, available LLMs can be fine-tuned on the recruiting domain and used as an additional scoring method, either in parallel or incorporated as one component in the Total Score ( $TS$ ).

However, this raises the challenge of keeping the tool computationally efficient, as SOTA LLMs are quite expensive to operate [22]. Incorporating feedback mechanisms for users is currently done in an offline fashion by utilizing their feedback on classification performance to re-calibrate the thresholds over time. This could also be done automatically at predefined intervals or by using distribution shift [19] or performance monitoring approaches [11].

## References

1. **Bilenko, M., Mooney, R. J. (2003).** On evaluation and training-set construction for duplicate detection. Proceedings of the KDD-2003 Workshop on Data Cleaning, Record Linkage, and Object Consolidation, pp. 7–12.
2. **Burk, H., Javed, F., Balaji, J. (2017).** Apollo: Near-duplicate detection for job ads in the online recruitment domain. 2017 IEEE International Conference on Data Mining Workshops, pp. 177–182. DOI: 10.1109/ICDMW.2017.29.
3. **Chan, L., Hogaboam, L., Cao, R. (2022).** Artificial intelligence in human resources. Springer International Publishing, pp. 139–155. DOI: 10.1007/978-3-031-05740-3\_9.
4. **Donovan, K. (2017).** Automated recruiting and the human factor.
5. **Draisbach, U. (2022).** Efficient duplicate detection and the impact of transitivity. Ph.D. thesis, Universität Potsdam. DOI: 10.25932/publishup-57214.
6. **Draisbach, U., Naumann, F. (2013).** On choosing thresholds for duplicate detection. Proceedings of the 18th International Conference on Information Quality.
7. **Ektefa, M., Jabar, M. A., Sidi, F., Memar, S., Ibrahim, H., Ramli, A. (2011).** A threshold-based similarity measure for duplicate detection. 2011 IEEE

- Conference on Open Systems, pp. 37–41. DOI: 10.1109/ICOS.2011.6079233.
8. **Engelbach, M., Klau, D., Scheerer, F., Drawehn, J., Kintz, M. (2023).** Fine-tuning and aligning question answering models for complex information extraction tasks. Accepted for KDIR 2023 - 15th International Conference on Knowledge Discovery and Information Retrieval. DOI: 10.48550/arXiv.2309.14805.
  9. **Foltýnek, T., Meuschke, N., Gipp, B. (2019).** Academic plagiarism detection: A systematic literature review. *ACM Computing Surveys*, Vol. 52, No. 6, pp. 1–42. DOI: 10.1145/3345317.
  10. **Gao, J., He, Y., Zhang, X., Xia, Y. (2017).** Duplicate short text detection based on Word2vec. 2017 8th IEEE International Conference on Software Engineering and Service Science, pp. 33–37. DOI: 10.1109/ICSESS.2017.8342858.
  11. **Ginart, T., Jinye-Zhang, M., Zou, J. (2022).** MLDemon: Deployment monitoring for machine learning systems. *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, Vol. 151, pp. 3962–3997.
  12. **Gunawan, D., Sembiring, C., Budiman, M. A. (2018).** The implementation of cosine similarity to calculate text relevance between two documents. *Journal of physics: conference series*, Vol. 978, pp. 012120. DOI: 10.1088/1742-6596/978/1/012120.
  13. **Kong, Y., Xie, C., Wang, J., Jones, H., Ding, H. (2021).** AI-assisted recruiting technologies: Tools, challenges, and opportunities. *Proceedings of the 39th ACM International Conference on Design of Communication*, pp. 359–361. DOI: 10.1145/3472714.3473697.
  14. **Leong, C. (2018).** Technology and recruiting 101: how it works and where it's going. *Strategic HR Review*, Vol. 17, No. 1, pp. 50–52. DOI: 10.1108/SHR-12-2017-0083.
  15. **Levenshtein, V. I. (1966).** Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, Vol. 10, No. 8, pp. 707–710.
  16. **Maier, D. (1978).** The complexity of some problems on subsequences and supersequences. *Journal of the ACM*, Vol. 25, No. 2, pp. 322–336. DOI: 10.1145/322063.322075.
  17. **Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013).** Efficient estimation of word representations in vector space. *Proceedings of the 1st International Conference on Learning Representations*.
  18. **Park, J., Lu, C., Marion, L. (2009).** Cataloging professionals in the digital environment: A content analysis of job descriptions. *Journal of the American Society for Information Science and Technology*, Vol. 60, No. 4, pp. 844–857. DOI: 10.1002/asi.21007.
  19. **Rabanser, S., Günemann, S., Lipton, Z. (2019).** Failing loudly: An empirical study of methods for detecting dataset shift. *Advances in Neural Information Processing Systems*, Vol. 32.
  20. **Ramya, R., Venugopal, K., Iyengar, S., Patnaik, L. (2016).** Feature extraction and duplicate detection for text mining: A survey. *Global Journal of Computer Science and Technology: C Software and Data Engineering*, Vol. 16, No. 05.
  21. **Reimers, N., Gurevych, I. (2020).** Making monolingual sentence embeddings multilingual using knowledge distillation. DOI: 10.48550/arXiv.2004.09813.
  22. **Samsi, S., Zhao, D., McDonald, J., Li, B., Michaleas, A., Jones, M., Bergeron, W., Kepner, J., Tiwari, D., Gadepally, V. (2023).** From words to watts: Benchmarking the energy costs of large language model inference. 2023 IEEE High Performance Extreme Computing Conference, pp. 1–9. DOI: 10.1109/HPEC58863.2023.10363447.

23. **Singhal, A. (2001).** Modern information retrieval: A brief overview. *IEEE Data Engineering Bulletin*, Vol. 24, No. 4, pp. 35–43.
24. **Sparck-Jones, K. (1972).** A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, Vol. 28, No. 1, pp. 11–21. DOI: 10.1108/eb026526.
25. **Sureka, A., Jalote, P. (2010).** Detecting duplicate bug report using character N-Gram-Based features. 2010 Asia Pacific Software Engineering Conference, pp. 366–374. DOI: 10.1109/APSEC.2010.49.
26. **Tian, Y., Sun, C., Lo, D. (2012).** Improved duplicate bug report identification. 2012 16th European Conference on Software Maintenance and Reengineering, pp. 385–390. DOI: 10.1109/CSMR.2012.48.
27. **Vo, M. T., Vo, A. H., Nguyen, T., Sharma, R., Le, T. (2021).** Dealing with the class imbalance problem in the detection of fake job descriptions. *Computers, Materials and Continua*, Vol. 68, No. 1, pp. 521–535. DOI: 10.32604/cmc.2021.015645.
28. **Wang, J., Dong, Y. (2020).** Measurement of text similarity: A survey. *Information*, Vol. 11, No. 9. DOI: 10.3390/info11090421.
29. **Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., Hernandez-Abrego, G., Yuan, S., Tar, C., Sung, Y. H., Strope, B., Kurzweil, R. (2019).** Multilingual universal sentence encoder for semantic retrieval. DOI: 10.48550/arXiv.1907.04307.
30. **Zhao, Y., Chen, H., Mason, C. M. (2022).** A framework for duplicate detection from online job postings. *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Association for Computing Machinery*, pp. 249–256. DOI: 10.1145/3486622.3493928.

*Article received on 12/04/2024; accepted on 22/07/2024.*

*\*Corresponding author is Matthias Engelbach.*