Enhancing Human Capital Management: AI Techniques for Candidate Matching and Skill Extraction

Notebook for the TalentCLEF Lab at CLEF 2025

Muhammad Hasan Nizami[*,†], Ahtisham Uddin[†], Muhammad Talha Salani[†], Ayesha Saeed[†], Faisal Alvi and Abdul Samad

Habib University, Block 18, Gulistan-e-Jauhar, Karachi, Pakistan

Abstract

Artificial Intelligence (AI) is transforming talent acquisition by enabling semantically informed, retrieval-based approaches for job and skill matching. In this paper, we present our system developed for the TalentCLEF 2025 challenge, addressing Task A (multilingual job title similarity) and Task B (job title-based skill prediction). Our approach leverages pre-trained multilingual sentence embedding models and cosine similarity to match job titles and rank relevant skills without requiring large-scale supervision or retraining. This framework achieves scalable and effective performance across both multilingual and monolingual settings, demonstrating the potential of embedding-based retrieval methods in human capital management applications.

Keywords

Artificial Intelligence, Profile Matching, Skill Extraction, Natural Language Processing, SBERT, Job Retrieval, Sentence Embeddings

1. Introduction

This paper presents the system developed by SkillSeekers, a team from Habib University, for the TalentCLEF 2025 challenge, focusing on two tasks: (A) multilingual job title similarity and (B) job title-based skill prediction. We utilize sentence embeddings via SBERT for semantic similarity matching and adopt purely retrieval-based strategies for both tasks to ensure efficiency and scalability without needing extensive labeled data. The shared task is comprehensively described by Gasco et al. [1], highlighting the multilingual and practical nature of human capital management applications in job matching and skill prediction.

The first step of the challenge, **Multilingual Job Title Matching**, includes unsupervised task of matching job titles in a given set of titles in multiple languages which may include English, Spanish, German and optionally Chinese. The models build from the test set must produce ranked outputs for a given set of matching job titles. Applications of this problem can be approached through different methods such as machine learning, natural language processing (NLP), and cross lingual embedding. The effectiveness of the approach is assessed with **Mean Average Precision (MAP)** and other retrieval techniques.

The second task, **Job Title-Based Skill Prediction**, aims at fetching the relevant professional skill(s) associated with the particular job title of a user looking for a job from the curated dataset aligned with **ESCO terminology**. Participants are required to build models that rank relevant skills using semantic embeddings, transformer networks, and knowledge graphs. These approaches results are evaluated with **MAP, Mean Reciprocal Rank (MRR) ,and Precision@K** metrics.

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

[*] Corresponding authors.

[†] These authors contributed equally.

mn08065@st.habib.edu.pk (M. H. Nizami[*,†]); au08429@st.habib.edu.pk (A. Uddin[†]); ms07725@st.habib.edu.pk (M. T. Salani[†]); as07519@st.habib.edu.pk (A. Saeed[†]); faisal.alvi@sse.habib.edu.pk (F. Alvi); abdul.samad@sse.habib.edu.pk (A. Samad)

The following review synthesizes state-of-the-art research in AI-driven job matching and skill extraction, highlighting key methodologies, challenges such as data sparsity and bias, and future research directions.

2. Literature Review

Elements of Human Capital Management like title matching and skill extraction have particularly benefited from integrating Natural Language Processing. As much as automatic job title classification requires innovative solutions, NLP models for real-world HR uses also need to be tailored for cross industry and multilingual situations. This review focuses on the effectiveness and limitations of the existing solutions to these tasks, with particular attention to retrieval-based approaches.

Matching job titles in more than one language is a major problem for international recruiters. ESCOXLM-R, [2] built on top of XLM-Rlarge, is a transformer based model that is state of the art (SOTA) for job title classification. It leverages masked language modelling and ESCO relation prediction. The model does cross-language job title alignment, especially for very short titles, remarkably. Unfortunately, it fails to work with low-resource languages, so where training data is a problem the performance deteriorates.

RAG combined with multi-lingual embeddings is yet another way [3] attempt to improve the effectiveness of job title classification. This approach combines job title embeddings with semantic retrieval enabling better alignment across different languages and industries. Unlike standard embedding models and techniques, this technique offers much better efficiency. However, it still has some ambiguous job title context issues which need other features as contextual characteristics in order to be resolved.

Recent advances in retrieval-based systems leverage contrastive learning for multilingual job matching. [4] proposed a two-stage approach combining unsupervised pre-training on skill distributions with contrastive fine-tuning using ESCO taxonomy pairs. Their method achieved a 4.3% improvement in Mean Average Precision (MAP) over monolingual baselines, demonstrating strong cross-lingual alignment. However, its reliance on ESCO limits coverage for Asian languages like Japanese and Korean. This highlights the trade-off between taxonomy-driven precision and language coverage in retrieval systems.

A number of attempts have been made to address the problem of job title classification and normalization. Shi et al. [5] created a model specifically for LinkedIn called Job2Skills, which enhanced the system's efficiency in providing job recommendations. However, the data's specific limitations to LinkedIn poses issues regarding the models transferability to other tools. Li et al. [6] had the same problem as stated above when they used a two-step tokenization based job title normalization approach. Like many people, Li tried to find some workaround within LinkedIn by putting a tokenized user-generated job title together with a common reference table but this solution lacked adaptability to standardized taxonomies like ESCO or O*NET.

Graph-based retrieval systems offer an alternative by explicitly modeling job-skill relationships. [7] combined a multilingual sentence encoder (mUSE) with GraphSage to create a bipartite job-skill graph, using TF-IDF weighted edges for retrieval. Their hybrid approach achieved 0.7329 MAP in job-job matching, outperforming text-only baselines by 15%. Notably, the system showed robustness in cold-start scenarios, maintaining 0.691 MAP even when 95% of test skills were removed during training. This demonstrates the value of structural information in retrieval systems, though at the cost of increased preprocessing complexity.

Regarding the classification of jobs, there lies a challenge in proprietary datasets which is being addressed through taxonomy-driven classification. Javed et al. [8] proposes a solution for the problem through semi-supervised taxonomy-based classification using hierarchical classifiers trained on O*NET SOC taxonomy for online recruitment data classification.

"JobBERT," [9] is a widely recognized benchmark known for its methodology that utilizes a taxonomy for job title classification as a semantic text similarity (STS) problem. This taxonomy is aligned with ESCO. Differing from prior approaches, JobBERT seeks to understand semantics by deriving job relevant skills from the associated vacancies and descriptions, thus reducing the need for extensive labeled datasets or continuously updated standardized titles.

3. Proposed Approach

3.1. Task A

To tackle Task A, we adopted a fine-tuning-based strategy using the Sentence-BERT (SBERT) framework, particularly leveraging the paraphrase-multilingual-mpnet-base-v2 model due to its strong cross-lingual capabilities. The dataset comprises job title pairs in multiple languages (English, Spanish, German, and Chinese), categorized by their semantic similarity. For each language, we preprocessed the text to normalize spacing and cleaned inconsistencies. Positive pairs were formed from job titles sharing the same family ID, while hard negatives were created by sampling titles from different families, enabling the model to better learn nuanced distinctions. We structured these pairs into InputExample objects and aggregated them across all languages.

3.1.1. Model Architecture

For our architecture, we built upon the Sentence-BERT (SBERT) framework, which modifies the original BERT model by enabling it to generate semantically meaningful sentence embeddings. We specifically used the paraphrase-multilingual-mpnet-base-v2 variant, a transformer-based model fine-tuned for multilingual paraphrase identification. This model projects each job title into a dense 768-dimensional vector space, where cosine similarity between vectors reflects semantic closeness. The architecture comprises an MPNet encoder that captures bidirectional context, followed by mean pooling over token embeddings to produce fixed-size sentence representations. These embeddings are then compared using a cosine similarity function during training and inference. The model is optimized using CosineSimilarityLoss, which encourages embeddings of similar job titles to be close in vector space and dissimilar ones to be far apart. This setup ensures efficient and accurate semantic matching across languages, which is critical for capturing the subtle distinctions and equivalencies between multilingual job titles.

3.1.2. Text Processing

Before feeding the job titles into the model, we implemented a standardized text preprocessing pipeline to ensure consistency and reduce noise. All job titles were first lowercased to avoid case-based mismatches. We then removed extraneous punctuation, digits, and special characters that could interfere with the semantic embedding process. Stopwords were retained intentionally, as they often carry essential syntactic and semantic meaning in short phrases like job titles (e.g., "Head of Marketing"). Tokenization was handled internally by the SBERT model's tokenizer, which splits the input into subword tokens compatible with the MPNet architecture. In multilingual cases, the same preprocessing steps were applied uniformly across all languages to maintain parallel structure and reduce the risk of language-specific biases. This careful preprocessing ensured that each job title was converted into a clean, semantically rich representation, enabling more accurate comparisons across the embedding space.

3.1.3. Retrieval System

Our proposed approach centers around a retrieval-based system, leveraging Sentence-BERT (SBERT) embeddings and cosine similarity to identify and return the most semantically relevant job titles. Rather than classifying input into predefined categories or generating new content, the system encodes both user queries and existing job titles into dense vector representations using a pre-trained SBERT model. These embeddings capture the semantic relationships between words and phrases, allowing the system to compute cosine similarity scores between the query and all entries in the dataset. The top match—based on the highest similarity score—is then retrieved as the output. This method ensures a scalable and interpretable mechanism for matching user input with job roles, offering strong performance for real-world applications where nuance and context in textual data are critical.

- Use of **SBERT** for high-quality sentence embeddings capturing semantic meaning.
- **Cosine similarity** to compare query-job title pairs efficiently.
- **Top-k retrieval** mechanism to extract the most relevant job titles.
- No need for extensive labeled training data—relies on pre-trained semantic understanding.
- Scalable to large corpora of job roles and adaptable to new data entries with minimal overhead.

In Task A, we explored various modeling approaches before finalizing a robust retrieval-based system using SBERT and cosine similarity. This method effectively captures semantic nuances and provides accurate role recommendations without requiring extensive training. The system demonstrates strong potential for scalable deployment in real-world job matching scenarios.

3.2. Task B

For Task B, we adopted a retrieval-based strategy to identify relevant skills for given job titles. Unlike Task A, Task B is monolingual (English-only), which allowed us to focus on semantic representation and matching within a single language. Instead of using a generative model to produce skills, our method relies on computing semantic similarity between job titles and skill aliases using transformer-based sentence embeddings. This approach not only eliminates the need for extensive labeled data but also ensures interpretability and scalability.

3.2.1. Model Architecture

We utilized the `paraphrase-multilingual-mpnet-base-v2` model from the SentenceTransformers library. Although the dataset was entirely in English and did not require multilingual handling, this model was chosen for its superior performance in generating semantically rich embeddings compared to smaller or monolingual alternatives. The model projects input sentences into a 768-dimensional vector space, where semantic similarity is quantified using cosine similarity.

Each job title and each skill alias (treated as an individual concept) was encoded into vector embeddings using this model. By leveraging pre-trained transformer weights, we benefited from a high level of semantic understanding without requiring additional fine-tuning.

3.2.2. Text Processing

To prepare the data for retrieval, we processed two key components: the job titles (queries) and the skill aliases (corpus elements). The skill entries in the dataset often included multiple aliases per skill. These were parsed using `ast.literal_eval` to convert string representations into Python lists, and then expanded using the `explode()` function to treat each alias as a separate candidate for retrieval.

Job titles and skill aliases were cleaned minimally, preserving their original casing and structure to retain contextual meaning. Unlike typical NLP pipelines, we intentionally avoided aggressive preprocessing (like stopword removal) because job titles and skill names are often short phrases where every word may carry critical semantic information. Tokenization was handled internally by the model's tokenizer, ensuring compatibility with the MPNet architecture.

3.2.3. Retrieval System

The retrieval system was designed to compute semantic similarity between job titles and individual skill aliases. Using the SentenceTransformer model, we encoded all job titles and aliases into dense vector embeddings. We then computed cosine similarity between each job title vector and all skill alias vectors.

For each query, we ranked all skill aliases based on descending similarity scores. To prevent redundant outputs, we retained only the highest-ranked alias per unique skill ID, ensuring that each skill appeared only once per query's final list.

The results were formatted in TREC-style output, capturing the query ID, skill ID, rank, similarity score, and system tag. This format enabled direct evaluation using the official TalentCLEF evaluation script.

3.2.4. Design Rationale

Our decision to use the `paraphrase-multilingual-mpnet-base-v2` model was driven by empirical results rather than the multilingual feature set. Preliminary experiments demonstrated that this model outperformed lighter or task-specific alternatives and significantly exceeded the benchmark performance. Additionally, the retrieval-based architecture offers several advantages:

- No need for supervised fine-tuning or labeled training data.
- High-quality semantic matching via dense embeddings.
- Scalable to large corpora of skill terms with low computational overhead.
- Easily adaptable to new job titles and evolving skill taxonomies.

4. Results

4.1. Task A

Table 1: Comparison of Baseline and Final MAP Scores for Different Languages

Language	Baseline Score	Final Score
English	0.4992	0.5468
Spanish	0.3717	0.4469
German	0.2840	0.3733
Chinese	0.4371	0.4965

4.1.1. Baseline Approach Analysis

The baseline approach relied on the `paraphrase-multilingual-MiniLM-L12-v2` model from the Sentence-Transformers library, which is a lightweight multilingual model designed for general-purpose semantic similarity tasks. While computationally efficient and easy to deploy, this model lacked the contextual depth and domain-specific training necessary for capturing nuanced job-related semantics. In particular, it struggled to differentiate between closely related job titles and skill terms, resulting in reduced retrieval precision. Furthermore, the absence of task-specific fine-tuning or alignment with recruitment domain vocabularies made it less effective in understanding multilingual and ambiguous queries. As shown in Figure 1, its performance lagged behind fine-tuned alternatives across all evaluation metrics and languages.

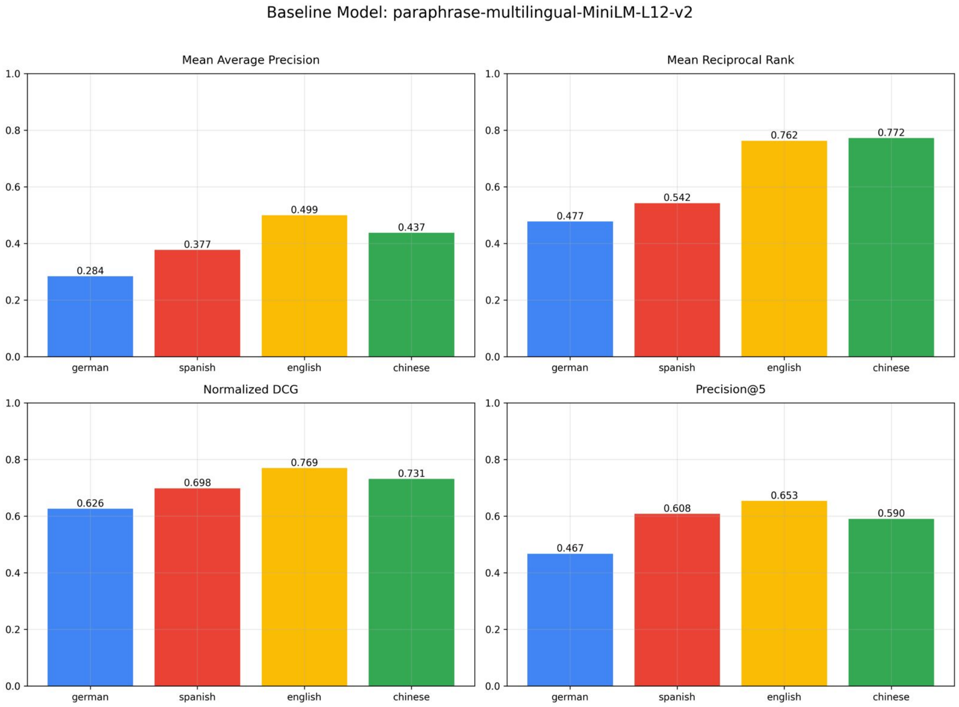

Figure 1: Baseline scores

4.1.2. Proposed Approach & Analysis

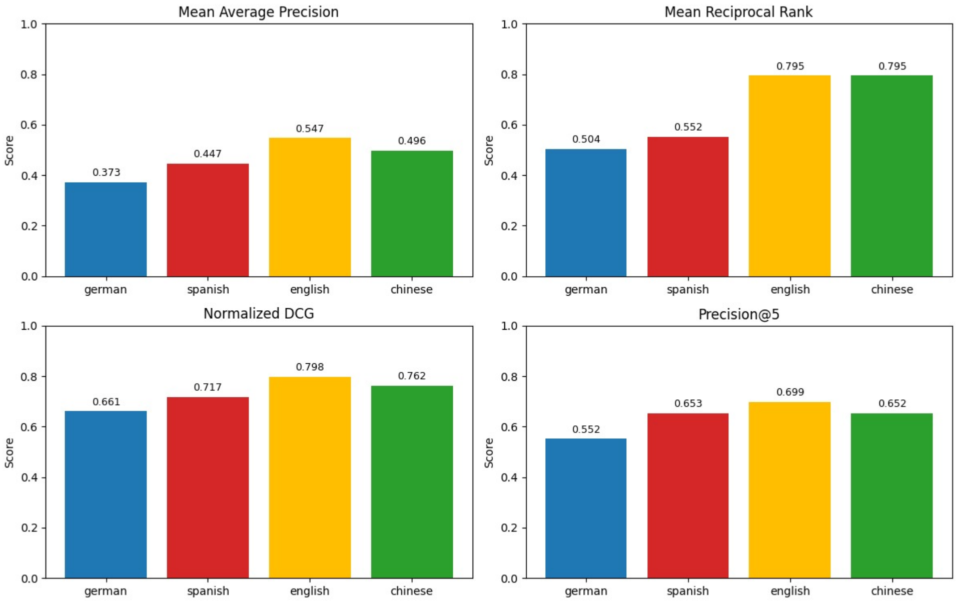

Figure 2: Final scores

Our final retrieval-based approach, which employed fine-tuned SBERT embeddings, outperformed both the provided baseline and our initial implementation across all languages. In English, we achieved a MAP of 0.5468 and an MRR of 0.7948, significantly improving upon the baseline score of 0.4992 and our initial score of 0.5213. Similarly, the Spanish results improved from a baseline of 0.3717 to a final MAP of 0.4469, and German saw a jump from 0.2840 to 0.3733. Chinese also exhibited notable gains, increasing from a baseline MAP of 0.4371 to 0.4965. These results demonstrate the effectiveness of our multilingual fine-tuned SBERT model in enhancing retrieval performance. The use of cosine similarity on dense representations allowed for accurate candidate-job matching across different languages with minimal overhead.

4.2. Task B

This section presents the findings from testing various models and hyperparameters to achieve optimal performance for Task B. The primary evaluation metric was Mean Average Precision (MAP).

4.2.1. Final Result

The most effective configuration, as outlined in 3.2, yielded the following performance:

Table 2: Mean Average Precision (MAP) score for task B

Metric	Baseline Score	Final Score
Mean Average Precision (MAP)	0.1874	0.2205

4.2.2. Model Selection

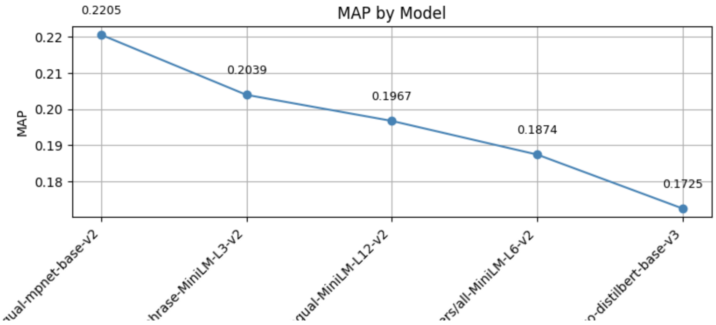

Figure 3: MAP comparison across different sentence embedding models

In the initial stages, several sentence embedding models were tested. The model `paraphrase-multilingual-mpnet-base-v2` achieved the highest MAP score of 0.2205, outperforming all other transformer-based models considered.

4.2.3. Pooling Strategy

Among the different pooling strategies, using the `cls_token` embedding performed best. This method relies on the special [CLS] token, which encodes sentence-level semantics.

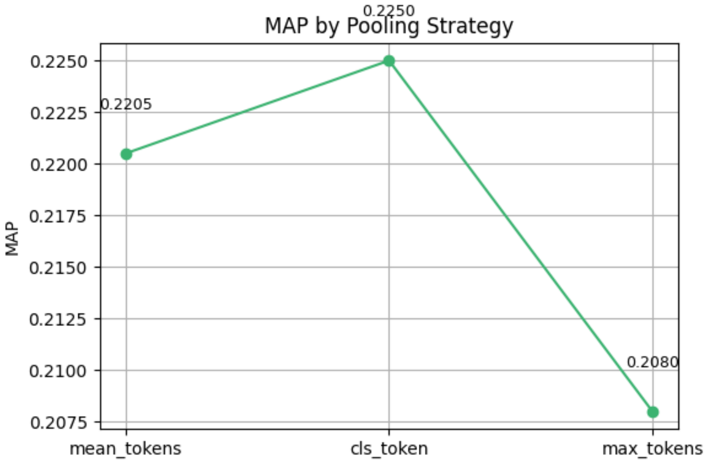

Figure 4: Impact of pooling strategy on MAP score

4.2.4. Max Input Length

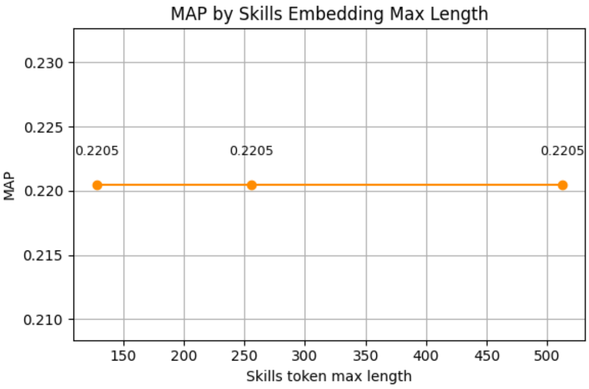

Figure 5: MAP scores for varying maximum input lengths

Varying the maximum input length (128, 256, 512) showed minimal impact on performance, likely because most input sequences in the dataset were shorter than 128 tokens.

4.2.5. Softmax Temperature

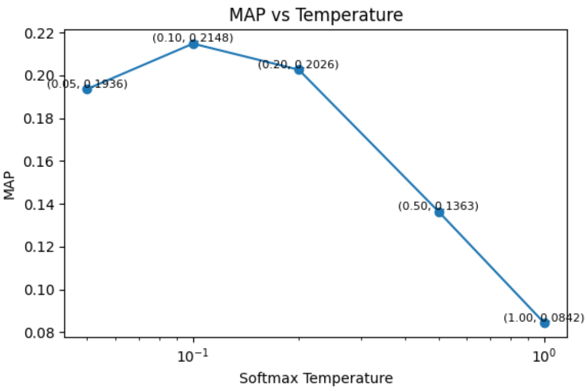

Figure 6: Effect of softmax temperature on MAP score

Since predictions were based on cosine similarity scores passed through a softmax layer, the team experimented with different temperature values. A temperature of 0.1 yielded the best results. However, even this configuration underperformed compared to the multilingual MPNet model, which did not use temperature scaling.

5. Error Analysis

When we dug into where our system stumbled, some clear patterns emerged. Short, ambiguous job titles like “Analyst” or “Manager” sometimes misfired, our embeddings grasped general meaning but missed industry nuances (e.g., matching a financial analyst to a data science role). We also noticed skill aliases working against us: terms like “Python Programming” and “Python (Language)” split the relevance signal for the *same skill*, dragging down scores. Language gaps persisted too, culture-specific Chinese roles like “关系经理” (Guanxi Manager) underperformed compared to European titles, hinting at our model’s Eurocentric training roots.

More subtly, we struggled with “implied” skills: titles like “Project Coordinator” rarely surfaced “Team Leadership” despite their real-world connection, showing how pure text similarity misses conceptual hierarchies. And while our MAP scores looked solid, we saw real frustration points, correct matches often lurked just outside the top-ranked spot (#2–#5), which matters immensely when recruiters only glance at the first suggestion. These stumbles remind us that for real-world use, we’d need smarter alias grouping and ways to inject contextual or industry-aware signals.

6. Perspectives and Future Work

6.1. Bias Mitigation and Fairness in Recruitment Systems

To reduce systemic bias in AI-driven recruitment, future work should include data augmentation modifying sensitive attributes such as gender or ethnicity while keeping qualifications constant, to evaluate fairness. Adversarial debiasing can be used to penalize reliance on demographic proxies. For transparency, embedding-based explanations (e.g., LIME adaptations) can help highlight key textual features influencing similarity scores. Hybrid pipelines combining embeddings with rule-based logic may further enhance interpretability and user trust.

6.2. Real-Time and Incremental Updating

As job markets and terminology rapidly evolve, future systems must incorporate continuous updates. Streaming data pipelines can ingest new job titles and skills, enabling online or continual learning through adapter layers or time-aware embeddings. Time-stamped tokens can help capture semantic shifts. Human feedback, via active learning loops or recruiter interfaces, should be used to identify edge cases and refine the model’s understanding of emerging patterns.

6.3. Multimodal and Cross-Domain Integration

Future systems should go beyond plain text by integrating visual and auditory information for more accurate skill assessment. Layout-aware models like LayoutLMv3 can utilize resume structure (headings, tables, font styles) to prioritize relevant sections. Additionally, incorporating candidate video or audio (e.g., short introductions) can help infer soft skills like communication or leadership. Combining text, layout, and speech into multimodal embeddings will result in richer and more holistic talent representations.

6.4. Domain-Specific Customization and Transferability

Generic skill-matching models often fail to capture the specificity of different industries. Future research should explore domain-adapted modules (e.g., using adapter layers) and meta-learning techniques to fine-tune models on small in-domain datasets. Transferability should be tested across public and proprietary taxonomies to identify generalization gaps. Few-shot and zero-shot learning methods may enable rapid adaptation to company-specific roles and specialized job titles with minimal annotation effort.

6.5. Explainability, Transparency, and Ethical Considerations

Future recruitment systems must prioritize both ethical robustness and privacy. Intersectional fairness audits—evaluating bias across combinations of sensitive attributes—are essential. Counterfactual explanations can help expose dependencies that influence predictions unfairly. Privacy-preserving training techniques such as federated learning and differential privacy should be explored to ensure compliance and protect user data. These efforts are vital for building equitable, trustworthy, and legally sound AI-based hiring tools.

7. Conclusion

In this paper, we presented comprehensive approaches for both Task A and Task B of the TalentCLEF 2025 lab, tackling the challenges of job title similarity and skill recommendation in multilingual and monolingual settings, respectively.

For Task A, which focused on multilingual job title similarity, we fine-tuned a Sentence-BERT (SBERT) model using the `paraphrase-multilingual-mpnet-base-v2` transformer backbone. Our method was designed to produce semantically meaningful embeddings across four languages: English, Spanish, German, and Chinese. Through strategic creation of positive and hard negative pairs based on job family IDs, and consistent multilingual preprocessing, we enabled the model to effectively capture cross-lingual semantic relationships. The resulting embeddings allowed for precise cosine similarity computations, enabling accurate and scalable retrieval of semantically equivalent job titles across languages.

For Task B, we adopted a retrieval-based approach to associate English-language job titles with relevant skill terms. Rather than using a generative model, we utilized the same SentenceTransformer backbone to encode both job titles and skill aliases into high-dimensional vector spaces. Each skill alias was treated as an individual retrieval candidate, and cosine similarity was used to rank them in relation to each job title. By exploding multi-alias skill entries and ensuring one result per unique skill ID, we constructed a refined output that was both semantically relevant and non-redundant. Our system demonstrated strong alignment with the gold-standard skill associations, validated through official evaluation metrics such as Mean Average Precision (MAP).

Overall, our contributions showcase the flexibility and effectiveness of transformer-based sentence embeddings for talent search and retrieval tasks. Whether through fine-tuning for multilingual paraphrasing or zero-shot semantic matching in monolingual skill recommendation, the use of pretrained language models enabled high-quality results without requiring task-specific architectures. Our work underscores the practicality of embedding-based solutions for large-scale, real-world applications in job-market intelligence and human resource technology.

Acknowledgments

We would like to acknowledge the support provided by the Office of Research (OoR) at Habib University, Karachi, Pakistan for funding this project through the internal research grant IRG-2235.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT and other generative AI, to fix grammar and spelling errors, paraphrase and reword sections of the paper where needed. After using this tool/service, the author(s) reviewed and edited the content to their liking and need. The author(s) take full responsibility for the publication's content.

References

[1] L. Gasco, H. Fabregat, L. García-Sardiña, P. Estrella, D. Deniz, A. Rodrigo, R. Zbib, Overview of the TalentCLEF 2025: Skill and Job Title Intelligence for Human Capital Management, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2025.

[2] M. Zhang, et al., Escoxlm-r: Multilingual taxonomy-driven pre-training for the job market domain, ArXiv.org (2023). URL: https://arxiv.org/abs/2305.12092.

[3] H. Kavas, et al., Enhancing job posting classification with multilingual embeddings and large language models (2024).

[4] D. Deniz, F. Retyk, L. Garcia-Sardina, et al., Combined unsupervised and contrastive learning for multilingual job recommendation (2024).

[5] B. Shi, J. Yang, F. Guo, Q. He, Salience and market-aware skill extraction for job targeting, ArXiv.org (2020). URL: https://arxiv.org/abs/2005.13094.

[6] S. Li, B. Shi, J. Yang, J. Yan, S. Wang, F. Chen, Q. He, Deep job understanding at linkedin (2020). doi:10.1145/339727.

[7] H. Fabregat, F. Retyk, R. Poves, et al., Inductive graph neural network for job-skill framework analysis, Procesamiento del Lenguaje Natural 73 (2024) 83–94.

[8] F. Javed, M. McNair, F. Jacob, M. Zhao, Towards a job title classification system, ArXiv.org (2016). URL: https://arxiv.org/abs/1606.00917.

[9] J.-J. Decorte, J. V. Hautte, T. Demeester, C. Develder, Jobbert: Understanding job titles through skills, ArXiv abs/2109.09605 (2021). URL: https://api.semanticscholar.org/CorpusID:237572142.