Article

Zero-Shot Recommendation AI Models for Efficient Job–Candidate Matching in Recruitment Process

Jarosław Kurek [1,2,*], Tomasz Latkowski [1], Michał Bukowski [1], Bartosz Świderski [1], Mateusz Łępicki [2], Grzegorz Baranik [2], Bogusz Nowak [2], Robert Zakowicz [1] and Łukasz Dobrakowski [2]

[1] Department of Artificial Intelligence, Institute of Information Technology, Warsaw University of Life Sciences, ul. Nowoursynowska 159, 02-776 Warsaw, Poland; tlatkowski@gmail.com (T.L.); michal.bukowski@buksoft.pl (M.B.); bartosz_swiderski@sggw.edu.pl (B.Ś.); robert_zakowicz@sggw.edu.pl (R.Z.)

[2] Avenga IT Professionals sp. z o.o., ul. Gwiaździsta 66, 53-413 Wroclaw, Poland; mateusz.lepicki@avenga.com (M.Ł.); grzegorz.baranik@avenga.com (G.B.); bogusz.nowak@avenga.com (B.N.); lukasz.dobrakowski@avenga.com (Ł.D.)

[*] Correspondence: jaroslaw_kurek@sggw.edu.pl

Abstract: In the evolving realities of recruitment, the precision of job–candidate matching is crucial. This study explores the application of Zero-Shot Recommendation AI Models to enhance this matching process. Utilizing advanced pretrained models such as all-MiniLM-L6-v2 and applying similarity metrics like dot product and cosine similarity, we assessed their effectiveness in aligning job descriptions with candidate profiles. Our evaluations, based on Top-K Accuracy across various rankings, revealed a notable enhancement in matching accuracy compared to conventional methods. Specifically, the all-MiniLM-L6-v2 model with a chunk length of 768 exhibited outstanding performance, achieving a remarkable Top-1 accuracy of 3.35%, 55.45% for Top-100, and an impressive 81.11% for Top-500, establishing it as a highly effective tool for recruitment processes. This paper presents an in-depth analysis of these models, providing insights into their potential applications in real-world recruitment scenarios. Our findings highlight the capability of Zero-Shot Learning to address the dynamic requirements of the job market, offering a scalable, efficient, and adaptable solution for job–candidate matching and setting new benchmarks in recruitment efficiency.

Keywords: recommendation model; pretrained nlp models; job–candidate matching; recruitment process

1. Introduction

The recruitment process, a critical component of human resource management, has undergone significant transformations in the digital age. Traditional methods of talent acquisition, often time-consuming and dependent on subjective human judgment, are increasingly being supplemented and, in some cases, replaced by artificial intelligence (AI)-driven solutions. This paper delves into an innovative aspect of this technological advancement: Zero-Shot Recommendation AI models, which promise to revolutionize the efficiency of job–candidate matching in recruitment.

This work represents a preliminary pilot approach to the problem, with fine-tuning and further refinement planned as the next stages of our research. By adopting a phase-wise methodology, we aim to iteratively enhance the model's accuracy and reliability, tailoring it more closely to the nuances of the recruitment domain.

Recruitment is an intricate process, balancing the need for speed with the necessity for accurate and fair candidate assessment. Traditional methods have struggled to keep pace with the rapidly changing job market, where new roles emerge continuously, and the skills required for existing roles evolve. Moreover, the sheer volume of job applications for positions, especially in populous sectors, can overwhelm human resources, leading to inefficiencies and potential biases in candidate selection. These challenges underscore the need for an intelligent, adaptive, and efficient recruitment process.

The recruitment process, encompassing the analysis of job descriptions, candidate CV screening, shortlisting, interviews, and job offers, can be partially automated with the aid of Artificial Intelligence (AI) solutions. These systems, known as Job Recommendation Systems (JRS) [1–3], are designed to support these processes.

Among the most prevalent types of JRS are Content-Based JRS (CB-JRS), Collaborative Filtering JRS (CF JRS), Hybrid JRS (H-JRS), and Knowledge-Based JRS (KB JRS) [4,5]. Each type has distinct characteristics and methodologies for recommending jobs to users:

- Content-Based JRS (CB-JRS): These systems recommend jobs to users based on the description of the jobs and a profile of the user's preferences. The recommendation is made by matching the content of the job postings, such as required skills, experience level, and job roles, with the user's profile, which includes their skills, past experiences, and preferences. This approach assumes that jobs similar to those a user liked in the past will also be of interest.
- Collaborative Filtering JRS (CF JRS): CF JRS recommend jobs based on the past behavior of users and similar decisions made by other users. This method does not require job content but relies on user–job interactions. The underlying assumption is that users who agreed in the past will agree in the future, and thus, jobs liked by similar users are recommended.
- Hybrid JRS (H-JRS): Hybrid systems combine both content-based and collaborative filtering approaches to leverage the advantages of both. By integrating these methodologies, H-JRS can provide more accurate and relevant job recommendations. This approach helps to mitigate the limitations of each system when used independently, such as the cold start problem in collaborative filtering and the limited content analysis in content-based systems.
- Knowledge-Based JRS (KB JRS): These systems recommend jobs based on explicit knowledge about users and jobs. Unlike other types, KB JRS do not solely rely on users' past behavior but use a knowledge base that includes rules and constraints to match users with suitable jobs. This approach is particularly useful when there is a need to consider complex requirements and preferences that are difficult to capture through user–job interactions alone.

Current Job Recommendation Systems (JRS) exhibit inherent limitations that significantly impact their effectiveness in dynamic recruitment environments. These limitations primarily stem from their reliance on historical data, leading to challenges in adapting to the rapidly evolving job market and the emergence of new roles. Traditional JRS models, such as Content-Based, Collaborative Filtering, and Hybrid systems, while useful, often fail to capture the nuanced requirements of new job positions and the diverse skill sets of candidates. This shortfall is particularly evident in scenarios where there is a lack of sufficient historical data for novel roles, resulting in inadequate or biased recommendations [1]. Furthermore, existing JRS models tend to suffer from data biases, which can skew recommendations and perpetuate existing inequities in the recruitment process. The lack of Explainable AI (XAI) in these systems further obscures the decision-making process, making it difficult to assess and improve the fairness and transparency of recommendations [6,7].

In anticipation of these challenges, significant enhancements are planned to be implemented in subsequent updates of both the ZSL models and those that have undergone fine-tuning. These enhancements are to include:

1. The integration of bias-reduction mechanisms and the development of methodologies aimed at significantly lowering biases, adhering to the principles of fair-aware modeling.
2. The incorporation of Explainable AI (XAI) to evaluate the model's ability to generalize, thereby enhancing the transparency and fairness of the recommendations.
3. The development of a specialized sub-model for use in the scoring and matching processes of recruitment, with the aim of ensuring the validation of research objectives and the achievement of key milestones.

These planned updates are aimed at refining the approach, ensuring that the recommendation system becomes more equitable, transparent, and responsive to the evolving needs of the job market.

These future directions are aimed at refining the ZSL approach, making it more robust, fair, and adaptable to the changing needs of the recruitment domain.

Current JRS types, despite their widespread use, are fraught with imperfections [8]. In the market, prevalent JRSs mainly support administrative processes and lack AI implementation. Globally, commercial solutions either support only a fraction of staffing services, do not incorporate AI, or have a severely limited scope in candidate search factors.

The integration of AI in recruitment has been a significant advancement, automating and refining various aspects of the process. AI-driven systems can handle large volumes of applications, screen resumes more efficiently than humans, and even assess candidate suitability through sophisticated algorithms. However, these systems often rely on historical data to function effectively. In scenarios where data are limited or non-existent, such as for newly created roles or unique skill sets, their effectiveness diminishes. This is where the concept of Zero-Shot Learning becomes pivotal.

Zero-Shot Learning, a recent breakthrough in AI, enables models to make accurate predictions or recommendations without having been explicitly trained on those specific tasks or categories [9–11]. In the context of recruitment, Zero-Shot Recommendation AI models can match candidates to jobs they have never encountered before. This capability is particularly valuable in today's dynamic job market, where emerging roles and skill sets can render traditional datasets obsolete. The ability of these models to generalize from learned information to new, unseen tasks holds the potential to significantly enhance the recruitment process, making it more agile and inclusive.

This paper aims to explore the development, capabilities, and implications of Zero-Shot Recommendation AI models in recruitment [12]. We will examine how these models are built, the algorithms that power them, and their practical applications in real-world recruitment scenarios. By assessing their performance against traditional AI models and human recruiters, we seek to demonstrate their potential in overcoming current challenges in talent acquisition. This research contributes to the evolving field of AI in recruitment, offering insights into how Zero-Shot Learning can lead to more efficient, unbiased, and adaptable recruitment processes.

With the advent of advanced AI technologies, the recruitment process is undergoing a significant transformation. The integration of AI-driven solutions, such as Zero-Shot Recommendation AI models, promises to revolutionize job–candidate matching, making it more efficient and unbiased. This paper explores these pretrained models and their potential to enhance the recruitment process.

Recent developments in e-recruitment have led to an increase in online job descriptions and a surge in job seekers submitting resumes, creating a massive pool of data [13]. This abundance of information necessitates efficient job–candidate matching processes, where recommender system technology plays a crucial role. Semantic technologies, in particular, have shown promise in improving e-recruitment by guiding document processing and automatic matching, thus enhancing job recommendation results [13].

Another approach involves using college-specific online job board systems, which aid students in finding suitable job opportunities [14–16]. These systems offer personalized job suggestions based on student abilities and assist businesses in candidate matching, demonstrating the potential of targeted job recommendation systems.

The concept of hybrid filtering in job recommendation systems has also been explored. By combining user and company datasets, these systems match user profiles with appropriate companies using various recommendation algorithms [17–20]. This approach, which includes content-based, collaborative, and hybrid filtering, addresses the limitations of individual methods, offering a more comprehensive solution.

However, a major concern in recruitment recommendation domains is the Matching Scarcity Problem (MaSP), where candidates or job vacancies suffer from a lack of matching

opportunities [21–24]. Strategies to identify and mitigate MaSP involve introducing changes in curricula and job descriptions to approximate candidates to semantically related jobs, thereby reducing the number of CVs and jobs suffering from MaSP.

In light of these developments, Zero-Shot Recommendation AI models emerge as a promising solution to the challenges faced in the recruitment process. These models, capable of making accurate predictions or recommendations without explicit training on specific tasks, are particularly valuable in scenarios with limited or non-existent data, such as newly created roles or unique skill sets [25]. The ability of Zero-Shot Learning to generalize from learned information to new, unseen tasks holds the potential to significantly enhance recruitment processes, making them more agile and inclusive.

Recent advancements in NLP and AI have led to the development of more sophisticated job recommendation systems. For instance, a study by Vijaya Kumari [26] explores the use of NLP in recommender systems, specifically in the domain of online recruitment. This approach involves analyzing resumes, profiles, and job descriptions to improve the matching process, addressing the "cold start" problem where new job postings and candidate profiles are not adequately matched.

Another innovative approach is the job recommendation method based on attention layer scoring characteristics and tensor decomposition, as proposed by Mao et al. [27]. This method focuses on users' attention levels and interactive behaviors, offering a more nuanced understanding of job seekers' interests and preferences.

Alsaif et al. [28] introduced a bi-directional recommendation system that supports both recruiters and job seekers. This system uses machine learning to process text content and similarity scores, enabling recruiters to find suitable candidates and job seekers to find matching job offers.

Furthermore, Dhameliya and Desai [29] proposed a hybrid job recommendation system combining content-based and collaborative filtering techniques. This system addresses the limitations of individual methods, offering a more comprehensive solution for e-recruitment.

A study by Zheng et al. [25] explores the use of generative models in job recommendations, demonstrating the potential of large language models in creating job descriptions from resumes. Khaire [30] provides a comprehensive review of resume analysis and job recommendation using AI, highlighting the role of machine learning and NLP in improving the recruitment process. Özcan [31] discusses the Classification Candidate Reciprocal Recommendation system, which uses classification techniques for matching candidates and job advertisements. Finally, Lamikanra and Obafemi-Ajayi [32] introduce the Beetle platform, leveraging AI and Blockchain technology to enhance the recruitment process.

These studies underscore the evolving recruitment ecosystem of AI in recruitment, where new technologies are being leveraged to improve the efficiency and effectiveness of job–candidate matching.

Integrating AI models with the Job Recommendation System (JRS) aims to significantly relieve recruiters at Avenga company from time-consuming and often repetitive tasks related to job analysis, large data sets analysis, crucial data extraction into databases, and candidate–job matching based on recommendations. These models are set to be implemented into the company's operations, where they will not only support the staffing service but, due to their unique nature, can also be utilized for other purposes, such as recruitment to the Avenga team.

The AI models will bolster innovative hybrid systems for recommending candidate profiles for specific queries and job positions for queries (Hybrid JRS). For the time being, zero-shot models constitute the initial stage of an innovative approach in the form of AI-assisted construction of the JRS system.

2. Materials and Methods

2.1. Dataset

The data source for constructing the various AI models comprises anonymized resumes in PDF and Word formats, obtained from potential candidates, as well as job specifications from Avenga. Currently, the resumes are limited to the English language only. Given the specific requirements of the models, both the resumes and job specifications have been segmented into chunks of lengths 512, 768, and 1024. These chunk lengths have been chosen for analysis to accommodate the models' processing capabilities and to ensure a comprehensive evaluation of their performance in matching candidates with job specifications.

Additionally, the dataset includes historical mappings of job-to-candidate pairings, which have been manually created by recruiters. It is important to note that not all resumes in the database, which were used to calculate the accuracy metric, have been reviewed by the recruiters. This lack of comprehensive review may lead to a significant overestimation of the error rate in the model's performance evaluations. The presence of manually curated job–candidate mappings serves as a valuable reference point; however, the potential for an inflated error due to unreviewed resumes must be taken into consideration when interpreting the results.

The dataset provides a comprehensive overview of the recruitment ecosystem, particularly emphasizing the IT sector and related technical fields. This focus reflects the dynamic nature of the tech industry, where the demand for specialized skills is both high and rapidly evolving. The predominance of IT-related job offers in our dataset highlights the sector's significant role in the current job market and underscores the dataset's utility in analyzing recruitment trends within this vital economic activity sector. Table 1 summarizes the key aspects of this analysis.

Table 1. Summary of job–candidate matching analysis.

Category	Count	Description
Unique Candidates	2432	Number of unique candidates in the dataset
Unique Jobs	1403	Number of unique jobs in the dataset
1 Job–1 Candidate	738	Number of matches where one job is matched with exactly one candidate
1 Job–Many Candidates	2431	Number of matches where one job is matched with multiple candidates
1 Candidate–Many Jobs	1246	Number of matches where one candidate is matched with multiple jobs
Total Records	3169	Total number of job–candidate matches in the dataset

The dataset includes a total of 2432 unique candidates and 1403 unique job positions. This diversity reflects the variety of specializations and roles in the contemporary job market. A significant portion of the dataset (738 instances) comprises scenarios where one job is matched with exactly one candidate. This one-to-one correspondence is indicative of highly specialized roles or unique candidate profiles that perfectly fit specific job descriptions.

On the other hand, there are 2431 instances where a single job is matched with multiple candidates. This scenario is common in job positions with broader or more general requirements, where a larger pool of candidates may possess the necessary qualifications or skills. Similarly, there are 1246 instances where a single candidate is matched with multiple job positions. This reflects the versatility of some candidates who may fit into various roles due to their diverse skill sets or adaptable profiles.

In total, the dataset contains 3169 records of job–candidate matches, providing a rich source for analysis and model training. The distribution of these matches (Table 2) offers insights into the dynamics of the job market and the versatility of the job-seeking population. This comprehensive dataset forms the basis of our subsequent analysis, wherein we apply Zero-Shot Recommendation AI models to enhance the efficiency and accuracy of job–candidate matching.

Table 2. Distribution of job offers by number of candidates.

Number of Candidates	Number of Jobs	Description
1	738	1 job–1 candidate
2	289	1 job–2 candidates
3	161	1 job–3 candidates
4	76	1 job–4 candidates
5	51	1 job–5 candidates
6	28	1 job–6 candidates
7	25	1 job–7 candidates
8	9	1 job–8 candidates
9	6	1 job–9 candidates
10	4	1 job–10 candidates
11	1	1 job–11 candidates
12	2	1 job–12 candidates
13	2	1 job–13 candidates
14	2	1 job–14 candidates
15	1	1 job–15 candidates
16	1	1 job–16 candidates
17	2	1 job–17 candidates
18	1	1 job–18 candidates
21	1	1 job–21 candidates
28	1	1 job–28 candidates
33	1	1 job–33 candidates
48	1	1 job–48 candidates
Total	1403	

2.2. Data Preprocessing

The preprocessing of curriculum vitae (CV) documents, initially in PDF format, for the application of pretrained NLP models involves several crucial steps. Initially, the text is extracted from PDF documents using a PDF-to-HTML converter, which inherently tags the text according to the formatting present in the original PDF. Subsequently, all HTML tags are stripped away, and the text is purified using the beautifulsoup library in Python, ensuring the removal of any residual HTML content.

Given the focus on English-language documents at this stage, an automatic language filtering process is applied to exclude non-English text. This is achieved using the "xlm-roberta-base-language-detection" model, ensuring the dataset's uniformity in language.

The cleaned text is then segmented into predefined lengths—512, 768, and 1024 characters—facilitating manageable processing by NLP models. These segments form the basis for embedding generation, utilizing the SentenceTransformer framework to produce vector representations. The mean embedding for each segment length is computed, encapsulating the contextual essence of the text.

For each predefined segment length, the embeddings are generated for both job descriptions and resumes, subsequently integrated into the dataset as new columns corresponding to each segment's mean embeddings. The enriched dataset, now containing these embeddings, is saved in a Parquet file format, optimized for efficient data storage and retrieval, thereby concluding the preprocessing phase.

2.3. The Concept of Zero-Shot Learning

Zero-Shot Learning (ZSL) is an innovative approach in the field of machine learning and artificial intelligence, where a model learns to accurately predict or recognize classes that it has not seen during the training phase. Unlike traditional machine learning models that require extensive labeled datasets for each category they need to identify, ZSL models leverage the understanding of relationships and attributes learned from seen classes to generalize to unseen ones.

The essence of ZSL lies in its ability to bridge the gap between the training data and real-world scenarios where not all classes can be anticipated or labeled in advance. This is achieved by using semantic representations, such as attribute vectors or textual descriptions, which provide a high-level understanding of both seen and unseen categories. By mapping input features to these semantic spaces, ZSL models can infer the characteristics of unseen classes based on their similarity to or differences from known classes.

In the context of job–candidate matching within the recruitment process, Zero-Shot Learning offers a promising solution to dynamically adapt to emerging job roles and diverse candidate profiles without the need for constant retraining. By understanding the semantic representation of job descriptions and candidate qualifications, ZSL models can effectively match candidates to jobs they have not been explicitly trained on, thus enhancing the efficiency and adaptability of the recruitment process.

The application of ZSL in recruitment not only streamlines the matching process but also mitigates the biases associated with traditional models that rely heavily on historical data. By focusing on the inherent attributes and capabilities required for a job and those possessed by a candidate, ZSL fosters a more meritocratic and inclusive approach to talent acquisition.

Overall, Zero-Shot Learning represents a significant advancement in the application of AI for recruitment, offering a flexible, efficient, and fair mechanism for job–candidate matching in an ever-evolving job market.

2.4. Zero-Shot Learning in AI Recruitment Models

Traditional machine learning models in recruitment systems rely heavily on historical data to train algorithms, which can accurately match candidates to job roles that the system has previously encountered. However, this approach falls short in addressing the emergence of new job titles and responsibilities, a common occurrence in fast-evolving sectors such as technology, digital marketing, and creative industries.

ZSL circumvents this limitation by leveraging semantic relationships between known and unknown categories. In recruitment, this means understanding the underlying skills, experiences, and qualifications required for a job, irrespective of the job title. For instance, a ZSL-based recruitment model can identify the suitability of a candidate for a "Data Scientist" position, even if it has not been explicitly trained on that job title, by understanding the semantic similarity between the job description and the candidate's profile in terms of skills like statistical analysis, machine learning, and data visualization.

The application of ZSL in recruitment AI models enhances their adaptability and scalability, enabling them to keep pace with the rapidly changing job landscape. It not only improves the efficiency and accuracy of candidate–job matching but also significantly reduces the need for frequent retraining of the AI models with new data. As a result, ZSL-equipped AI recruitment models are better positioned to meet the needs of both employers and job seekers, providing a more dynamic, inclusive, and forward-looking approach to talent acquisition.

2.5. Pretrained Models for Zero-Shot Learning

Pretrained models are the foundation for the Zero-Shot Recommendation AI Models explored in this study [33–36]. Pretrained models, having been trained on vast and diverse datasets, encapsulate a broad understanding of natural language, making them ideal for Zero-Shot Learning (ZSL) tasks. The essence of ZSL in this context is to leverage these models to understand and match job descriptions with candidate profiles without explicit prior training on these specific tasks.

2.6. Transformer-Based Architectures

The core of our pretrained models lies in transformer-based architectures, which have revolutionized Natural Language Processing (NLP). Models such as BERT (Bidirectional Encoder Representations from Transformers), GPT (Generative Pretrained Transformer),

and their derivatives have shown exceptional capability in understanding and generating human-like text. Their ability to capture contextual relationships between words in a sentence has led to significant advancements in various NLP tasks.

2.7. Model Selection for Zero-Shot Learning

For the purpose of this study, we focused on models (Table 3) that exhibit strong performance in sentence and paragraph embedding, crucial for encapsulating the essence of job descriptions and candidate profiles. The selection criteria were based on the models' ability to produce embedding vector representations that accurately capture semantic meanings.

Table 3. Comparison of pretrained models for Zero-Shot Learning.

Model	Size (MB)	Parameters	Layers	Languages
LaBSE	1150	470 M	24	Multilingual
all-MiniLM-L12-v2	200	110 M	12	Multilingual
all-MiniLM-L6-v2	100	22 M	6	Multilingual
all-mpnet-base-v2	450	110 M	12	Multilingual
distiluse-base-multilingual-cased-v2	250	135 M	6	Multilingual
msmarco-distilbert-cos-v5	300	65 M	6	English
msmarco-distilbert-dot-v5	300	65 M	6	English
paraphrase-multilingual-MiniLM-L12-v2	220	85 M	12	Multilingual
paraphrase-multilingual-mpnet-base-v2	470	110 M	12	Multilingual

2.8. Criteria for Selecting Pretrained Models for Zero-Shot Learning

In the context of determining embeddings and calculating accuracy through either the dot product or cosine similarity, the choice of appropriate models for comparison with the basic model all-MiniLM-L6-v2 model is crucial for our paper. The following are pretrained models which had taken into consideration in this paper that could provide insightful comparisons:

1. all-MiniLM-L12-v2: As an extended version of the selected model with more layers (L12), it would be interesting to compare how the change in depth affects the quality of embeddings and the resulting accuracies.
2. paraphrase-multilingual-MiniLM-L12-v2: Trained on multilingual data, this model could offer interesting comparisons, especially if your data are also multilingual.
3. paraphrase-multilingual-mpnet-base-v2: Based on the MPNet architecture, this model could provide unique insights in comparison to the MiniLM architecture.
4. distiluse-base-multilingual-cased-v2: This simplified model could be interesting to compare in terms of performance and accuracy relative to more complex models.
5. all-mpnet-base-v2: Another model based on MPNet, which could provide comparisons with different MPNet variants.
6. msmarco-distilbert-cos-v5 and msmarco-distilbert-dot-v5: Both models are based on DistilBERT but differ in their approach to calculating similarity (cosine vs. dot product). Comparing them with MiniLM could provide insights into the efficiency of different similarity calculation methods.
7. LaBSE: Based on a large multilingual database, this could be interesting to compare with MiniLM, especially in the context of handling different languages.

The selection of these models should provide a good diversity in terms of architecture, training approach, and multilingual support, which could be key for in-depth analysis and comparison in our scientific article.

2.9. The Language-Agnostic BERT Sentence Embedding (LaBSE) Model

The Language-Agnostic BERT Sentence Embedding (LaBSE) [37–39] model represents a significant advancement in the field of Natural Language Processing (NLP), particularly in the realm of creating language-agnostic sentence embeddings. Developed and

released by Google, LaBSE is designed to efficiently produce high-quality embeddings for sentences across multiple languages, making it an invaluable tool in globalized and multilingual applications.

2.9.1. Model Architecture and Training

LaBSE builds upon the architecture of BERT (Bidirectional Encoder Representations from Transformers), a transformer-based model known for its effectiveness in various NLP tasks. The model consists of 24 transformer layers, with an embedding size of 1024, making it a large and powerful model capable of capturing complex language nuances.

The training process of LaBSE involved leveraging dual-encoder frameworks, where the model was trained on parallel corpora covering 109 languages. This extensive training has enabled the model to understand and encode sentences in a wide range of languages into a shared embedding space.

2.9.2. Features and Capabilities

One of the key features of LaBSE is its ability to maintain cross-lingual semantic accuracy. This means that semantically similar sentences in different languages are mapped close to each other in the embedding space, facilitating effective cross-lingual transfer and applications such as multilingual semantic search and text classification.

Moreover, LaBSE demonstrates superior performance in bilingual sentence-matching tasks and in sentence retrieval tasks across a variety of languages. This makes it particularly suitable for applications where understanding and comparing sentence-level semantic information across different languages is crucial.

2.9.3. Relevance to Zero-Shot Learning

In the context of Zero-Shot Learning (ZSL), LaBSE's language-agnostic capabilities allow it to perform effectively in scenarios where training data for certain languages or specific tasks is scarce or unavailable. By understanding the semantic relationships in a language-independent manner, LaBSE can be utilized to make accurate inferences about unseen data, making it an ideal choice for ZSL applications.

2.9.4. Application in Job-Candidate Matching

In the domain of job–candidate matching, LaBSE's proficiency in handling multilingual content is particularly advantageous. For global companies and recruitment processes involving diverse linguistic backgrounds, LaBSE can effectively process and compare job descriptions and candidate profiles in various languages. This facilitates a more inclusive and comprehensive matching process, aligning with the needs of a globalized job market.

In summary, the Language-Agnostic BERT Sentence Embedding model stands as a powerful tool in the arsenal of NLP, enabling efficient and accurate sentence embeddings across multiple languages. Its application extends to various domains, including but not limited to cross-lingual semantic searches, text classification, and, in the context of this study, enhanced job–candidate matching in recruitment processes.

2.10. The MiniLM Model

The MiniLM model [40–42] stands as a remarkable innovation in the realm of Natural Language Processing (NLP), particularly in the efficient creation of language models. MiniLM, short for Miniaturized Language Model, is designed with the primary goal of reducing the size and complexity of traditional transformer-based models while retaining their high performance.

2.10.1. Model Architecture and Training

MiniLM, developed by Microsoft, leverages the transformer architecture, similar to BERT and GPT, but with significant modifications to reduce its size. The model is

characterized by a smaller number of parameters and layers, making it more efficient in terms of computational resources and faster in inference time.

The training of MiniLM involves a distillation process, where knowledge from larger models like BERT or GPT-3 is transferred to MiniLM. This process involves training MiniLM to replicate the output of these larger models, enabling it to achieve similar levels of understanding and accuracy but with significantly reduced computational requirements.

2.10.2. Features and Capabilities

Despite its smaller size, MiniLM demonstrates remarkable capabilities in understanding and generating human-like text. It excels in various NLP tasks, including text classification, question answering, and sentence embedding. The model's efficiency does not come at the cost of performance, as it often matches or even exceeds the capabilities of its larger counterparts in certain applications.

A key feature of MiniLM is its versatility. Being lightweight, it can be easily integrated into various applications, especially those requiring real-time processing or operating under limited computational resources.

2.10.3. Relevance to Zero-Shot Learning

In the context of Zero-Shot Learning (ZSL), MiniLM's efficiency and robust performance make it particularly suitable. Its ability to generalize well from learned data to unseen tasks is crucial in ZSL scenarios, where models are expected to make accurate predictions without explicit prior training on specific categories or tasks.

2.10.4. Application in Job-Candidate Matching

For job–candidate matching, MiniLM's compact size and fast processing capabilities enable it to quickly analyze and compare large volumes of text data, such as resumes and job descriptions. Its effectiveness in sentence embedding and text classification can be leveraged to accurately match candidates to job openings, enhancing the efficiency and scalability of the recruitment process.

In conclusion, the MiniLM model emerges as a powerful yet efficient tool in NLP, facilitating high-performance language processing tasks in a compact framework. Its application extends across various domains, including real-time language processing, efficient sentence embedding, and enhancing job–candidate matching in recruitment, making it a valuable asset in AI-driven recruitment systems.

2.11. The MPNet Model

MPNet [43–45], or Masked and Permuted Pre-training for Language Understanding, is an innovative model in the field of Natural Language Processing (NLP) that extends the capabilities of traditional transformer-based models. Developed by Microsoft, MPNet introduces a novel pre-training method that enhances the model's understanding of language context and structure.

2.11.1. Model Architecture and Training

MPNet builds upon the transformer architecture, similar to models like BERT, but introduces a unique pre-training strategy that combines aspects of both masked language modeling (MLM) and permuted language modeling (PLM). This hybrid approach allows MPNet to capture a more comprehensive understanding of sentence structure and context.

During its training, MPNet learns to predict masked tokens in a sentence not only based on the surrounding context (as in MLM) but also taking into account the sentence's overall sequence information (as in PLM). This dual approach enables MPNet to develop a more nuanced understanding of language, leading to improved performance in various downstream NLP tasks.

2.11.2. Features and Capabilities

MPNet is particularly noted for its effectiveness in tasks that require a deep understanding of sentence context and structure, such as sentence embedding, semantic similarity assessment, and text classification. Its ability to accurately model the relationships between words in a sentence contributes to its superior performance in these areas.

Another key feature of MPNet is its efficiency. Despite its complex pre-training strategy, the model remains computationally efficient and can be applied in various real-world applications, including those with constraints on processing power and inference time.

2.11.3. Relevance to Zero-Shot Learning

In the realm of Zero-Shot Learning (ZSL), MPNet's advanced understanding of language context and structure makes it an ideal candidate. Its ability to generalize from the training data to unseen scenarios is crucial for ZSL, where the model must make accurate predictions on tasks or categories not encountered during training.

2.11.4. Application in Job-Candidate Matching

MPNet's strength in understanding complex language structures makes it well-suited for analyzing and comparing the text data involved in job–candidate matching. Its proficiency in semantic understanding can be leveraged to accurately align candidates' qualifications with job requirements, thus enhancing the efficiency and effectiveness of the recruitment process.

In summary, MPNet stands out as a powerful model in NLP, combining the strengths of MLM and PLM to offer enhanced language understanding. Its application in various domains, including job–candidate matching and Zero-Shot Learning, showcases its versatility and effectiveness in handling complex language processing tasks.

2.12. The DistilBERT Model

DistilBERT [46,47], short for Distilled Bidirectional Encoder Representations from Transformers, represents a significant stride in the field of Natural Language Processing (NLP). Developed as a streamlined and optimized version of the BERT model, DistilBERT offers a balance between performance and efficiency, making it highly suitable for various NLP applications.

2.12.1. Model Architecture and Training

DistilBERT is a smaller and faster version of the original BERT model, retaining much of its predecessor's performance capabilities while significantly reducing its size and computational requirements. This is achieved through a process known as knowledge distillation, where the DistilBERT model is trained to replicate the behavior of the larger BERT model.

During training, DistilBERT learns from the "teacher" BERT model by mimicking its output on a large corpus of text. This approach allows DistilBERT to capture the essence of BERT's language understanding capabilities but in a more compact form, resulting in a model that is approximately 40% smaller and 60% faster.

2.12.2. Features and Capabilities

Despite its reduced size, DistilBERT exhibits strong performance across a range of NLP tasks, including text classification, sentiment analysis, and question answering. Its efficiency makes it particularly appealing for applications where model size and speed are crucial, such as on mobile devices or in environments with limited computational resources.

A key advantage of DistilBERT is its versatility. The model can be easily integrated into existing systems that originally applied BERT, allowing for a seamless transition to a more efficient framework without significant loss in performance.

2.12.3. Relevance to Zero-Shot Learning

In Zero-Shot Learning (ZSL) scenarios, DistilBERT's ability to provide high-quality language representations with reduced complexity is invaluable.

2.13. Adaptation for Zero-Shot Learning

The adaptation of these pretrained models for ZSL involves fine-tuning the models to align with the specific context of job–candidate matching. The fine-tuning process adjusts the models' weights to optimize them for generating embeddings that reflect the similarity between job descriptions and candidate profiles, enabling the Zero-Shot Recommendation AI Models to effectively match candidates to jobs they have not been explicitly trained on.

The utilization of these advanced pretrained models provides a robust foundation for the development of Zero-Shot Recommendation AI Models in recruitment. Their ability to understand and process complex language patterns enables these models to perform accurate job–candidate matching, transforming the recruitment process to be more efficient and inclusive.

2.14. Similarity Metrics

In the context of our analysis for job–candidate matching, two primary similarity metrics were applied to evaluate the alignment between job descriptions and candidate profiles: the dot product and cosine similarity. These metrics are pivotal in quantifying the degree of similarity, thereby facilitating the effective pairing of job openings with suitable candidates.

2.14.1. Dot Product

The dot product, also known as the scalar product, is a fundamental operation in vector algebra. Given two vectors $\mathbf{A} = [a_1, a_2, ..., a_n]$ and $\mathbf{B} = [b_1, b_2, ..., b_n]$, their dot product is defined as follows :

$$\mathbf{A} \cdot \mathbf{B} = \sum_{i=1}^{n} a_i b_i = a_1 b_1 + a_2 b_2 + ... + a_n b_n \tag{1}$$

In our analysis, each job description and candidate profile is represented as a vector in a high-dimensional space, where each dimension corresponds to a feature derived from the textual data. The dot product between a job vector and a candidate vector yields a scalar value that reflects the degree of similarity between them. A higher dot product indicates a greater degree of similarity, suggesting a stronger match between the job requirements and the candidate's qualifications.

2.14.2. Cosine Similarity

Cosine similarity is another widely used metric in text analysis and information retrieval. It measures the cosine of the angle between two vectors, providing an indication of their orientation with respect to one another, regardless of their magnitude. The cosine similarity between two vectors $\mathbf{A}$ and $\mathbf{B}$ is given by

$$\text{cosine\_similarity}(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} \tag{2}$$

where $\|\mathbf{A}\|$ and $\|\mathbf{B}\|$ are the Euclidean norms (or magnitudes) of the vectors $\mathbf{A}$ and $\mathbf{B}$, respectively. The cosine similarity ranges from -1 to 1, where 1 indicates identical orientation, 0 indicates orthogonality (no similarity), and -1 indicates opposite orientation.

In the context of our study, cosine similarity provides a measure of how closely a candidate's profile aligns with the job description in terms of direction in the feature space, irrespective of the length of the vectors. This is particularly useful for comparing profiles and job descriptions of varying lengths and detail, ensuring that the similarity measure is not biased by the quantity of text but rather by the relevance of the content.

Both the dot product and cosine similarity are integral to our methodology, enabling the quantitative assessment of job–candidate matches. These metrics facilitate a nuanced analysis of the dataset, allowing us to identify the most suitable candidates for each job position based on the textual data provided in their profiles and job descriptions.

2.15. Comparison of Dot Product and Cosine Similarity

While both dot product and cosine similarity are effective in measuring the similarity between two vectors, they have distinct characteristics that make them suitable for different scenarios. This subsection outlines the advantages and disadvantages of each metric, providing insights into their applicability in various contexts.

2.15.1. Advantages

Dot Product

- Simplicity: The dot product is straightforward to compute, requiring only the summation of the products of corresponding elements.
- Scalability: It scales linearly with the dimensionality of the vectors, making it computationally efficient for high-dimensional data.
- Sensitivity to Magnitude: The dot product takes into account the magnitude of vectors, making it useful in applications where the size of the vectors is meaningful.

Cosine Similarity

- Normalization: Cosine similarity normalizes for vector length, focusing purely on the directionality of vectors. This makes it ideal for comparing documents of different lengths.
- Robustness: It is less sensitive to variations in magnitude, reducing the impact of outliers or large values that might skew the results.
- Intuitive Interpretation: The range of cosine similarity from -1 to 1 provides an intuitive scale for measuring similarity, where 1 represents identical directionality.

2.15.2. Disadvantages

Dot Product

- Magnitude Dependence: The sensitivity to the magnitude of vectors can be a drawback when the size difference between vectors (such as document length) should not influence the similarity measure.
- Lack of Normalization: Without normalization, comparisons across different sets or scales may be less meaningful.

Cosine Similarity

- Ignorance of Magnitude: While often an advantage, the disregard for vector magnitude means that cosine similarity cannot distinguish between vectors that are similar in direction but different in scale.
- Computational Complexity: The need to calculate vector norms for normalization can introduce additional computational overhead, especially for large datasets.

In summary, the choice between dot product and cosine similarity should be guided by the specific requirements of the application. The dot product may be preferable when the magnitude of the vectors carries important information, while cosine similarity is more suitable for cases where only the orientation or the pattern of the data matters, such as in text similarity measures where document length should not affect the similarity score.

2.16. Evaluation Metric: Top-K Accuracy

In the framework of our investigation into job–candidate matching utilizing Zero-Shot Recommendation AI Models, we implement the "Top-K Accuracy" metric as a crucial measure of performance. This metric plays a key role in determining the model's ability to accurately rank candidates based on their relevance to a specific job position.

Top-K Accuracy is a fundamental metric in recommendation systems and information retrieval fields. It measures the model's effectiveness in identifying the true candidate for a

job within the top K candidates, ranked according to their relevance scores. These relevance scores can be derived from cosine similarity or dot product calculations between the job description and the candidate profiles. Both metrics serve to quantify the alignment in a high-dimensional feature space, albeit through slightly different mathematical approaches.

The process to compute Top-K Accuracy for each job in our dataset involves several steps:

1. Rank candidates based on their relevance scores to the job description, with scores calculated using either cosine similarity or the dot product method, from highest to lowest.
2. Check whether the true candidate—the one who was actually hired or is deemed the most suitable for the job—is among the top K-ranked candidates.
3. A correct prediction is noted if the true candidate is within this top K group.

Accordingly, Top-K Accuracy is determined by the ratio of correct predictions to the total number of jobs assessed. This can be formulated as

$$\text{Top-K Accuracy} = \frac{\text{Number of Jobs with True Candidate in Top K}}{\text{Total Number of Jobs}} \tag{3}$$

Applying Top-K Accuracy in our analysis provides insight into the model's practical effectiveness in recruitment contexts. High Top-K Accuracy signifies the model's proficiency in narrowing down the pool of candidates, thereby facilitating the recruitment process and increasing the probability of successful job–candidate pairings. This metric is especially pertinent for evaluating the model's precision in initial candidate screenings within large applicant pools, where the ability to effectively shortlist candidates is essential.

3. Results and Discussion

This study's core objective was to evaluate the performance of various embedding models for job–candidate matching using Zero-Shot Recommendation AI Models. The efficacy of these models was assessed using the Top-K Accuracy metric across different rankings. This section discusses the outcomes of these evaluations as presented in Tables 4 and 5, and compares them with the baseline performance of a random model as shown in Table 6.

3.1. Analysis of Embedding Models Using Dot Product

The analysis of embedding models using the dot product metric, as detailed in Table 4, provides insightful conclusions regarding their performance in job–candidate matching. The values in the table represent percentages, indicating the accuracy of each model across various Top-K rankings.

The 'all-MiniLM-L6-v2 model', featuring vector lengths of 512, achieved the highest observed Top-1 accuracy among the models evaluated, reaching 3.35%. While this marks a notable performance, it's important to contextualize this achievement within the broader landscape of Job-Recommender Systems (JRSs) approaches. The Top-1 accuracy figure, though the highest in this comparison, underscores the challenges inherent in accurately matching job descriptions with candidate profiles on the first attempt. This performance, while promising, suggests there is significant room for improvement and exploration of alternative or complementary approaches to enhance accuracy further. Moreover, the model's performance trend across higher K values, culminating in a Top-500 accuracy of 77.55%, reflects its ability to provide a wider range of potentially suitable candidate matches as more options are considered. This pattern of increasing accuracy with higher K values indicates the model's utility in casting a wider net for candidate–job matches, although it also highlights the inherent trade-offs between precision and recall in such systems. The results underline the importance of continued research and development in this field to advance the capabilities of JRS technologies.

Table 4. Performance comparison of various embedding models using dot product across different Top-K rankings.

Model	Len.	Top1	Top50	Top100	Top150	Top200	Top250	Top300	Top350	Top400	Top450	Top500
LaBSE	512	1.71	26.23	36.64	44.83	50.25	55.31	58.45	61.65	64.58	67.36	68.78
LaBSE	768	2.00	26.16	38.13	45.69	50.32	54.31	57.45	60.01	63.15	65.07	67.78
LaBSE	1024	1.85	23.88	33.86	42.20	48.33	51.60	55.52	58.37	60.66	62.79	65.07
all-MiniLM-L12-v2	512	2.99	35.71	46.19	52.96	58.30	62.01	64.93	67.78	70.49	72.70	74.13
all-MiniLM-L12-v2	768	2.78	35.42	45.19	51.89	56.59	60.44	62.87	65.93	68.78	71.13	73.34
all-MiniLM-L12-v2	1024	3.14	33.50	45.40	51.39	55.74	60.37	63.58	66.14	69.14	71.35	73.20
all-MiniLM-L6-v2	512	2.85	37.28	49.25	56.45	61.15	65.15	68.57	71.35	74.13	75.91	77.55
all-MiniLM-L6-v2	768	3.28	36.92	51.03	58.23	63.36	67.50	69.57	71.49	74.48	75.69	77.41
all-MiniLM-L6-v2	1024	3.35	35.85	47.83	55.17	61.01	64.72	67.50	70.21	72.42	74.84	76.84
all-mpnet-base-v2	512	2.42	31.86	44.19	52.67	59.09	63.15	66.50	68.78	70.71	72.27	73.84
all-mpnet-base-v2	768	1.78	32.86	45.12	51.75	57.73	61.94	65.00	67.57	69.85	71.77	73.77
all-mpnet-base-v2	1024	2.99	31.72	45.47	53.96	59.02	62.15	65.00	68.00	70.06	71.70	73.70
distiluse-base-multilingual-cased-v2	512	2.42	33.64	44.12	50.82	56.66	61.44	65.00	68.57	71.49	73.84	76.13
distiluse-base-multilingual-cased-v2	768	2.14	31.22	41.48	49.68	55.10	60.09	63.72	66.50	69.00	71.28	73.13
distiluse-base-multilingual-cased-v2	1024	2.00	28.51	40.63	48.75	53.31	56.95	60.30	63.36	66.36	68.28	70.63
msmarco-distilbert-cos-v5	512	1.71	29.94	40.91	49.96	55.31	58.45	63.22	66.22	69.71	71.63	72.77
msmarco-distilbert-cos-v5	768	2.49	32.57	42.77	49.75	55.60	59.52	62.58	65.72	68.42	70.35	71.92
msmarco-distilbert-cos-v5	1024	2.21	29.22	41.20	48.97	54.31	58.59	62.87	65.65	68.71	70.49	72.77
msmarco-distilbert-dot-v5	512	1.85	31.36	41.63	50.11	55.52	60.44	63.44	66.07	68.42	70.42	73.20
msmarco-distilbert-dot-v5	768	3.06	30.93	42.05	49.18	54.31	59.02	62.44	65.15	68.57	70.92	72.27
msmarco-distilbert-dot-v5	1024	2.99	30.29	41.05	47.40	51.96	55.60	59.94	62.58	65.00	67.93	70.35
paraphrase-multilingual-MiniLM-L12-v2	512	1.78	25.23	38.28	46.04	51.32	55.67	59.94	62.94	65.43	67.50	69.71
paraphrase-multilingual-MiniLM-L12-v2	768	1.78	23.95	35.21	42.34	48.61	52.67	57.66	61.87	64.50	67.28	69.42
paraphrase-multilingual-MiniLM-L12-v2	1024	1.35	24.80	35.85	42.84	48.18	53.17	56.45	59.52	62.01	65.57	67.57
paraphrase-multilingual-mpnet-base-v2	512	1.50	26.30	37.42	44.48	50.32	55.31	59.59	63.29	66.50	68.42	70.14
paraphrase-multilingual-mpnet-base-v2	768	1.92	26.30	36.78	43.69	49.18	55.17	59.44	62.30	65.36	67.64	70.28
paraphrase-multilingual-mpnet-base-v2	1024	2.28	27.23	37.56	45.05	50.89	55.31	58.66	61.44	64.15	66.57	68.85

Table 5. Performance comparison of various embedding models using cosine similarity across different Top-K rankings.

Model	Len.	Top1	Top50	Top100	Top150	Top200	Top250	Top300	Top350	Top400	Top450	Top500
LaBSE	512	1.85	26.66	38.28	45.90	52.67	56.66	59.52	63.36	66.07	68.21	70.49
LaBSE	768	1.92	27.58	36.92	44.83	51.75	56.59	60.51	63.44	65.72	67.78	69.21
LaBSE	1024	1.85	26.02	35.42	42.55	48.61	54.60	58.30	61.01	63.79	65.36	67.43
all-MiniLM-L12-v2	512	2.71	38.77	51.10	58.30	63.72	67.14	70.21	72.42	74.70	76.62	78.55
all-MiniLM-L12-v2	768	3.21	37.35	50.25	57.31	63.08	67.07	70.35	72.13	74.20	75.91	77.76
all-MiniLM-L12-v2	1024	3.42	37.06	49.68	57.45	62.37	66.57	69.78	72.42	73.70	75.84	77.33
all-MiniLM-L6-v2	512	3.35	40.56	53.96	61.44	67.14	71.92	74.55	76.48	78.33	79.47	80.97
all-MiniLM-L6-v2	768	3.42	40.48	55.45	62.72	67.36	71.13	74.41	76.76	78.55	79.69	81.11
all-MiniLM-L6-v2	1024	3.28	40.13	52.96	60.66	66.29	69.99	72.49	75.12	76.91	78.90	79.69
all-mpnet-base-v2	512	2.99	36.64	48.75	57.59	63.44	66.93	70.35	72.42	74.06	75.12	76.98
all-mpnet-base-v2	768	2.92	36.42	48.40	55.81	61.37	65.72	69.71	72.42	74.13	75.12	76.84
all-mpnet-base-v2	1024	3.56	36.85	49.04	57.16	62.08	66.00	68.85	71.49	73.63	74.98	76.27
distiluse-base-multilingual-cased-v2	512	2.92	33.78	46.76	55.02	60.51	64.58	68.00	70.28	73.34	75.77	77.05
distiluse-base-multilingual-cased-v2	768	2.92	32.50	46.04	53.96	59.66	64.01	67.28	69.78	72.06	74.06	75.27
distiluse-base-multilingual-cased-v2	1024	2.28	31.36	43.12	50.53	56.81	60.58	64.43	67.36	69.92	72.49	73.70
msmarco-distilbert-cos-v5	512	2.49	30.51	42.91	51.32	57.31	62.37	64.72	67.28	69.92	71.42	73.41
msmarco-distilbert-cos-v5	768	2.64	32.93	43.91	50.46	57.52	61.87	64.58	67.00	69.35	70.49	72.56
msmarco-distilbert-cos-v5	1024	2.57	30.43	42.98	51.60	58.30	62.30	65.36	68.07	69.71	71.56	73.41
msmarco-distilbert-dot-v5	512	2.42	29.51	41.13	47.68	52.46	57.09	60.80	63.36	65.93	68.42	70.78
msmarco-distilbert-dot-v5	768	2.28	28.44	39.49	46.97	52.89	56.45	60.09	62.87	65.36	68.64	71.13
msmarco-distilbert-dot-v5	1024	2.71	28.23	38.70	46.61	51.75	56.09	59.80	62.37	64.65	67.36	69.57
paraphrase-multilingual-MiniLM-L12-v2	512	2.00	31.58	44.62	53.24	58.73	62.79	66.50	69.00	71.20	73.06	74.91
paraphrase-multilingual-MiniLM-L12-v2	768	2.42	28.23	41.34	49.25	54.74	59.87	63.36	66.71	69.00	71.42	73.13
paraphrase-multilingual-MiniLM-L12-v2	1024	1.92	29.08	40.91	49.18	53.60	58.16	62.01	65.36	68.14	70.85	72.77
paraphrase-multilingual-mpnet-base-v2	512	2.21	32.29	45.33	53.10	58.52	62.51	66.50	69.07	71.20	72.77	75.05
paraphrase-multilingual-mpnet-base-v2	768	2.71	31.36	42.91	50.68	58.23	62.08	64.36	67.00	70.92	73.06	74.70
paraphrase-multilingual-mpnet-base-v2	1024	2.14	32.07	43.26	51.39	57.02	61.30	64.72	67.36	70.28	71.63	72.92

Table 6. Top-K Accuracy percentage of random baseline model across various rankings.

Top1	Top50	Top100	Top150	Top200	Top250	Top300	Top350	Top400	Top450	Top500
0.04	2.06	4.11	6.17	8.22	10.28	12.34	14.39	16.45	18.5	20.56

Conversely, the 'paraphrase-multilingual-MiniLM-L12-v2' model, especially with a vector length of 1024, showed the lowest Top-1 accuracy, at only 1.35%. While its performance improves with higher K values, it reaches 69.71% at Top-500, which is still lower compared to other models. This model's relatively lower accuracy across all K values suggests it is less effective for this specific application of job–candidate matching.

In general, as the Top-K value increases, all models demonstrate an improvement in accuracy. However, the 'all-MiniLM-L6-v2' model stands out for its consistently high performance, making it the preferred choice for practical applications in recruitment. In contrast, the 'paraphrase-multilingual-MiniLM-L12-v2' model, particularly at higher vector lengths, underperforms in comparison, indicating it may be less suited for scenarios requiring high precision in candidate matching.

3.2. Analysis of Embedding Models Using Cosine Similarity

This subsection presents a detailed analysis based on the performance of various embedding models using the cosine similarity metric, as summarized in Table 5. In this table, all values represent percentages, indicating the accuracy of each model in matching candidates to job descriptions for the top 1, top 50, up to top 500 candidates.

The 'all-MiniLM-L6-v2' model, with a vector length of 512, consistently shows the highest Top-1 accuracy across all models. This indicates its superior ability to accurately match the most relevant candidate to a job description on the first try. Notably, this model also maintains its lead in performance as the number of top candidates increases, demonstrating robustness in identifying suitable candidates within a broader pool. The high accuracy rates persist up to the top 500 candidates, peaking at 81.11%, which is significantly higher than the other models.

On the other end of the spectrum, the model with the lowest performance in the Top-1 category is the 'paraphrase-multilingual-MiniLM-L12-v2' with a vector length of 1024, showing an accuracy of only 1.92%. However, it is important to note that the performance of all models generally improves as the number of top candidates considered increases. Even the lower-performing models show a substantial increase in accuracy at higher K values, though they do not surpass the leading models.

In summary, the 'all-MiniLM-L6-v2' model with a vector length of 512 emerges as the most effective model for job–candidate matching using the cosine similarity metric. This model not only excels at identifying the most relevant candidate but also maintains superior performance across a wider range of top candidates. The results emphasize the importance of choosing the right model and vector length for efficient and accurate job–candidate matching in recruitment processes.

3.3. Comparison with Random Baseline Model

This subsection provides a comparative analysis between the random baseline model (as described in Table 6) and the pretrained models evaluated using cosine similarity and dot product. The random baseline model serves as a benchmark to assess the effectiveness of the embedding models in the task of job–candidate matching.

Below is a comparative analysis between the random baseline model and the embedding models evaluated using cosine similarity and dot product for specific Top-K accuracies (Top 1, Top 100, Top 200, Top 300, Top 400, and Top 500). The random baseline model, which selects candidates randomly, serves as a benchmark to assess the effectiveness of the embedding models in the task of job–candidate matching.

Top 1 Accuracy: The random baseline model has a Top-1 accuracy of 0.04%, as shown in Table 6. In contrast, the best-performing model in Tables 4 and 5, 'all-MiniLM-L6-v2'

with a vector length of 512, achieved a Top-1 accuracy of 3.35%. This stark difference highlights the superior capability of the 'all-MiniLM-L6-v2' model in accurately identifying the most suitable candidate for a job position on the first attempt.

Top 100 Accuracy: For Top 100 accuracy, the random baseline model's performance is significantly lower, with an accuracy of 4.11%. Meanwhile, the 'all-MiniLM-L6-v2' model demonstrates a much higher accuracy, reaching 55.45% when using cosine similarity, indicating its effectiveness in narrowing down the pool of candidates to the top 100 most relevant ones.

Top 200 Accuracy: The random model achieves a Top-200 accuracy of 8.22%, whereas the 'all-MiniLM-L6-v2' model significantly outperforms this with an accuracy of 67.36% (using cosine similarity), showcasing its ability to effectively identify a broader pool of suitable candidates.

Top 300 to Top 500 Accuracies: As we extend the comparison to Top 300, Top 400, and Top 500 accuracies, the random baseline model's accuracies are 12.34%, 16.45%, and 20.56% respectively. In comparison, the 'all-MiniLM-L6-v2' model maintains a high level of accuracy, achieving 74.41% for Top 300, 76.76% for Top 400, and 81.11% for Top 500 when using cosine similarity. This demonstrates the model's consistent performance even when considering a large number of top candidates.

In conclusion, the comparison with the random baseline model underscores the significant improvement in job–candidate matching accuracy achieved by the 'all-MiniLM-L6-v2' model. The substantial differences in accuracy for Top 1, Top 100, Top 200, Top 300, Top 400, and Top 500 highlight the model's effectiveness in accurately matching candidates to job descriptions, far surpassing the random selection approach.

3.4. Analysis of Model Performance

The analysis of the performance of various embedding models, as presented in Tables 4 and 5, offers insightful perspectives on their capabilities in the context of Zero-Shot Learning for job–candidate matching. This subsection aims to juxtapose these results against the intrinsic properties of the models discussed in the subsection 2.7, providing a comprehensive understanding of how specific model characteristics influence their performance in this domain.

3.4.1. Performance Evaluation Using Dot Product

Table 4 presents the performance of the selected models using the dot product as the similarity metric. The dot product, emphasizing the magnitude and directionality of the vectors, reflects how well the models capture the semantic essence of job descriptions and candidate profiles. Notably, models like 'all-MiniLM-L6-v2' demonstrated superior performance across various Top-K accuracies. This can be attributed to its efficient transformer architecture and the ability to generate dense embeddings that encapsulate the relevant features for the task at hand. The compact and optimized nature of 'MiniLM' variants, as discussed in the model selection subsection, evidently contributes to their effectiveness in this application, balancing computational efficiency with semantic precision.

3.4.2. Performance Evaluation Using Cosine Similarity

Conversely, Table 5 evaluates the models based on cosine similarity, which focuses on the orientation of the vectors rather than their magnitude. This metric is particularly useful in assessing the models' ability to identify the semantic alignment between job and candidate descriptions, irrespective of the length of the textual data. In this analysis, the 'all-MiniLM-L6-v2' model consistently outperforms others, reaffirming its robustness in capturing the semantic similarities crucial for accurate job–candidate matching. The high performance of models applying the 'MiniLM' architecture, which is designed for efficient NLP tasks with limited computational resources, underscores the significance of architectural choices in model efficacy.

3.4.3. Implications of Model Properties

The comparative analysis of model performance using dot product and cosine similarity metrics illuminates the impact of specific model properties, such as architectural efficiency, embedding quality, and computational optimization, on their suitability for Zero-Shot Learning in recruitment. The 'MiniLM' models, with their distilled architecture, stand out for their ability to generate high-quality embeddings with fewer parameters, making them particularly adept at this task. On the other hand, models like 'LaBSE' and 'MPNet', despite their broader linguistic coverage and complex architectures, may not offer the same level of optimization for this specific application, as reflected in their relatively lower performance metrics.

In conclusion, the results underscore the importance of model selection in Zero-Shot Learning applications within the recruitment domain. The analysis not only reaffirms the capabilities of transformer-based models like 'MiniLM' in handling semantic similarity tasks but also highlights the critical role of model properties, such as architectural design and embedding efficiency, in determining their effectiveness for job–candidate matching.

3.5. Understanding the Margin of Error in Model Performance

The performance of various embedding models for job–candidate matching, as quantified by metrics such as Top-1 accuracy, reveals inherent limitations that warrant a deeper investigation. Despite the advancements in Zero-Shot Recommendation AI Models, no model achieves an accuracy higher than 3.56% (model all-mpnet-base-v2 for 1024 chunk length) for Top-1 accuracy. This margin of error prompts an examination of the multifaceted challenges within the candidate selection process.

The low accuracies observed in our pilot approach are indicative of the complex nature of manual recruitment processes, not limited to the company from which data was sourced but reflective of broader industry practices. Key factors contributing to this margin of error include the ground truth dilemma, where recruiters are incapable of reviewing all CVs in a database, and human factors such as recruiter fatigue, varying skill levels, and subjective approaches to candidate evaluation. These elements underscore the discrepancies between model predictions and the ground truth, further compounded by the volume of work, time constraints, and the manual nature of candidate assessment.

Moreover, potential semantic underrepresentation of candidates' professional competencies and job offers, coupled with the challenges of polysemy in automated matching, highlight the intricate relationship between language use in CVs and job descriptions and their interpretation by AI models. The corpus of pre-trained models, while extensive, may not fully encapsulate the domain-specific language and subtleties present in recruitment materials, suggesting a need for model fine-tuning and domain adaptation.

In light of these insights, our subsequent approach will encompass fine-tuning of models to better align with the nuanced requirements of the recruitment domain. By adapting model parameters to more closely reflect the intricacies of job descriptions and candidate profiles, we aim to mitigate the identified limitations and enhance model accuracy in predicting suitable job–candidate matches.

4. Conclusions

In this paper, we explored the potential of Zero-Shot Recommendation AI Models in revolutionizing the recruitment process through efficient job–candidate matching. Our investigation delved into various pretrained NLP models and similarity metrics, notably the dot product and cosine similarity, to evaluate their effectiveness in this context. The results demonstrated a significant enhancement in matching accuracy and efficiency, highlighting the capability of these models to adapt to the dynamic demands of the job market.

The 'all-MiniLM-L6-v2' model, in particular, showed exceptional performance across different Top-K accuracy metrics, underscoring its robustness in identifying suitable candidates from a vast pool. This indicates a promising direction for leveraging compact

and efficient NLP models in recruitment, balancing computational resources with high-quality embeddings.

Furthermore, the use of Zero-Shot Learning techniques addresses the challenge of ever-evolving job roles and requirements, offering a scalable solution that does not rely on extensive labeled datasets. This not only increases the adaptability of recruitment processes but also promotes a more inclusive approach by reducing potential biases associated with traditional methods.

However, it is important to acknowledge the limitations and areas for further research. The adaptation of these models to specific recruitment contexts, the integration of diverse data sources, and the mitigation of any residual biases are crucial steps toward fully realizing the potential of AI in recruitment. Future work could also explore the combination of these AI models with human expertise to create a hybrid approach that maximizes the strengths of both.

In conclusion, Zero-Shot Recommendation AI Models hold significant promise for enhancing the recruitment ecosystem. By harnessing the power of advanced NLP and AI techniques, we can achieve a more efficient, fair, and adaptable recruitment process, ultimately leading to better job–candidate matches and contributing to the overall success of organizations and job seekers alike.

References

1. Patil, A.; Suwalka, D.; Kumar, A.; Rai, G.; Saha, J. A Survey on Artificial Intelligence (AI) based Job Recommendation Systems. In Proceedings of the 2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), Erode, India, 23–25 March 2023; pp. 730–737. [CrossRef]
2. Thali, R.; Mayekar, S.; More, S.; Barhate, S.; Selvan, S. Survey on Job Recommendation Systems using Machine Learning. In Proceedings of the 2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA), Coimbatore, India, 10–11. January 2024; pp. 453–457. [CrossRef]
3. Tayade, T.; Akarte, R.; Sorte, G.; Tayade, R.; Khodke, P. Data Mining Approach to Job Recommendation Systems. In Proceedings of the International Conference on Mobile Computing and Sustainable Informatics, Virtual Event, 17–19 September 2021; pp. 503–509.
4. De Ruijt, C.; Bhulai, S. Job Recommender Systems: A Review. *arXiv* **2021**, arXiv:2111.13576.
5. Freire, M.N.; de Castro, L.N. e-Recruitment recommender systems: A systematic review. *Knowl. Inf. Syst.* **2021**, *63*, 1–20. [CrossRef]
6. Kamble, A.; Tambe, S.; Bansode, H.; Joshi, S.; Raut, S. Job Recommendation System for Daily Paid Workers using Machine Learning. *Int. J. Res. Appl. Sci. Eng. Technol.* **2023**, *11*, 3086–3089. [CrossRef]
7. Shi, X.; Wei, Q.; Chen, G. A bilateral heterogeneous graph model for interpretable job recommendation considering both reciprocity and competition. *Int. J. Comput. Sci. Eng.* **2024**, *11*, 128–142. [CrossRef]
8. Delecraz, S.; Eltarr, L.; Becuwe, M.; Bouxin, H.; Boutin, N.; Oullier, O. Responsible Artificial Intelligence in Human Resources Technology: An innovative inclusive and fair by design matching algorithm for job recruitment purposes. *J. Responsible Technol.* **2022**, *11*, 100041. [CrossRef]
9. Xian, Y.; Lampert, C.H.; Schiele, B.; Akata, Z. Zero-Shot Learning—A Comprehensive Evaluation of the Good, the Bad and the Ugly. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 2251–2265. [CrossRef] [PubMed]
10. Romera-Paredes, B.; Torr, P.H.S., An Embarrassingly Simple Approach to Zero-Shot Learning. In *Visual Attributes*; Feris, R.S., Lampert, C., Parikh, D., Eds.; Springer: Berlin/Heidelberg, Germany, 2017; pp. 11–30. [CrossRef]

11. Socher, R.; Ganjoo, M.; Sridhar, H.; Bastani, O.; Manning, C.D.; Ng, A.Y. Zero-Shot Learning Through Cross-Modal Transfer. *Adv. Neural Inf. Process. Syst.* **2013**, *26*, 1–10.

12. Wang, J.; Krishnan, A.; Sundaram, H.; Li, Y. Pre-trained Neural Recommenders: A Transferable Zero-Shot Framework for Recommendation Systems. *arXiv* **2023**, arXiv:2309.01188.

13. Brek, A.; Boufaïda, Z. Semantic Approaches Survey for Job Recommender Systems. *DBLP* **2022**, *1*, 1–10.

14. Elavarasi, K.; Sivaranjani, S.; ArunKumar, M.; Santhosh, M. Recommendation System for Job Opportunities based on Candidate Parameters. In Proceedings of the 2023 Second International Conference on Electronics and Renewable Systems (ICEARS), Tuticorin, India, 2–4 March 2023; pp. 1504–1509. [CrossRef]

15. Mansourvar, M.; Mohd Yasin, N. Development of a Job Web Portal to Improve Education Quality. *Int. J. Comput. Theory Eng.* **2014**, *6*, 43–46. [CrossRef]

16. Parry, E.; Tyson, S. An analysis of the use and success of online recruitment methods in the UK. *Hum. Resour. Manag. J.* **2008**, *18*, 257–274. [CrossRef]

17. Mulay, A.; Sutar, S.; Patel, J.; Chhabria, A.; Mumbaikar, S. Job Recommendation System Using Hybrid Filtering. In Proceedings of the International Conference on Automation, Computing and Communication 2022, Navi Mumbai, India, 7–8 April 2022. [CrossRef]

18. Yadalam, T.V.; Gowda, V.M.; Kumar, V.S.; Girish, D.; M., N. Career Recommendation Systems using Content based Filtering. In Proceedings of the 2020 5th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 10–12 June 2020; pp. 660–665. [CrossRef]

19. Shah, K.; Salunke, A.; Dongare, S.; Antala, K. Recommender systems: An overview of different approaches to recommendations. In Proceedings of the 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), Coimbatore, India, 17–18 March 2017; pp. 1–4. [CrossRef]

20. Priyanka, P. A Survey Paper on Various Algorithm's based Recommender System. *IOSR J. Comput. Eng.* **2017**, *19*, 27–32. [CrossRef]

21. Cardoso, A.; Mourão, F.; Rocha, L. Mitigating Matching Scarcity in Recruitment Recommendation Domains. In Proceedings of the ACM, Virtual Event, 27–29 October 2020. [CrossRef]

22. Bian, S.; Zhao, W.X.; Song, Y.; Zhang, T.; Wen, J.R. Domain Adaptation for Person-Job Fit with Transferable Deep Global Match Network. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 4810–4820. [CrossRef]

23. Cardoso, A.; Mourão, F.; Rocha, L. A characterization methodology for candidates and recruiters interaction in online recruitment services. In Proceedings of the 25th Brazillian Symposium on Multimedia and the Web, Rio de Janeiro, Brazil, 29 October– 1 November 2019; pp. 333–340. [CrossRef]

24. Yan, R.; Le, R.; Song, Y.; Zhang, T.; Zhang, X.; Zhao, D. Interview Choice Reveals Your Preference on the Market: To Improve Job-Resume Matching through Profiling Memories. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 914–922. [CrossRef]

25. Zheng, Z.; Qiu, Z.; Hu, X.; Wu, L.; Zhu, H.; Xiong, H. Generative Job Recommendations with Large Language Model. *arXiv* **2023**, arXiv:2307.02157.

26. Kumari, S.V. Job Recommendation System Using NLP. *Int. J. Eng. Sci.* **2023**, *11*, 2721–2727. [CrossRef]

27. Mao, Y.; Cheng, Y.; Shi, C. A Job Recommendation Method Based on Attention Layer Scoring Characteristics and Tensor Decomposition. *Appl. Sci.* **2023**, *13*, 9464. [CrossRef]

28. Alsaif, S.; Hidri, M.S.; Ferjani, I.; Eleraky, H.A.; Hidri, A. NLP-Based Bi-Directional Recommendation System: Towards Recommending Jobs to Job Seekers and Resumes to Recruiters. *Big Data Cogn. Comput.* **2022**, *6*, 147. [CrossRef]

29. Dhameliya, J.; DESAI, N.P. Job Recommendation System using Content and Collaborative Filtering based Techniques. *Int. J. Soft Comput. Eng. (IJSCE)* **2019**, *1387*, 1. [CrossRef]

30. Khaire, S. Review on Resume Analysis and Job Recommendation using AI. *IJRASET* **2021**, *9*, 1221–1224. [CrossRef]

31. Ozcan, G.; Oguducu, S.G. Applying Classifications Techniques in Job Recommendation System for Matching of Candidates and Advertisements. *Int. J. Intell. Comput. Res.* **2017**, *8*, 798–806. [CrossRef]

32. Lamikanra, K.; Obafemi-Ajayi, T. Leveraging the Power of Artificial Intelligence and Blockchain in Recruitment using Beetle Platform. In Proceedings of the 2023 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics, Bengaluru, India, 27–28 January 2023; pp. 117–124.

33. Tiong, A.M.H.; Li, J.; Li, B.; Savarese, S.; Hoi, S.C.H. Plug-and-Play VQA: Zero-shot VQA by Conjoining Large Pretrained Models with Zero Training. *arXiv* **2022**, arXiv:2210.08773.

34. Phan, T.; Vo, K.; Le, D.; Doretto, G.; Adjeroh, D.; Le, N. ZEETAD: Adapting Pretrained Vision-Language Model for Zero-Shot End-to-End Temporal Action Detection. *arXiv* **2023**, arXiv:2311.00729.

35. Öztürk, E.; Ferreira, F.; Jomaa, H.S.; Schmidt-Thieme, L.; Grabocka, J.; Hutter, F. Zero-Shot AutoML with Pretrained Models. *arXiv* **2022**, arXiv:2206.08476.

36. Kang, H.; Blevins, T.; Zettlemoyer, L. Translate to Disambiguate: Zero-shot Multilingual Word Sense Disambiguation with Pretrained Language Models. *arXiv* **2023**, arXiv:2304.13803.

37. Galatolo., F.; Martino., G.; Cimino., M.; Tommasi., C. Dense Information Retrieval on a Latin Digital Library via LaBSE and LatinBERT Embeddings. In Proceedings of the 12th International Conference on Data Science, Technology and Applications-DATA. INSTICC, SciTePress, Rome, Italy, 11–13 July 2023; pp. 518–523. [CrossRef]

38. Gayathri, G.L.; Swaminathan, K.; Divyasri, K.; Durairaj, T.,; Bharathi, B. PANDAS@Abusive Comment Detection in Tamil Code-Mixed Data Using Custom Embeddings with LaBSE. In Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages, Dublin, Ireland, 26 May 2022; pp. 112–119. [CrossRef]

39. Han, L.; Sorokina, I.; Erofeev, G.; Gladkoff, S. cushLEPOR: Customising hLEPOR metric using Optuna for higher agreement with human judgments or pre-trained language model LaBSE. *arXiv* **2021**, arXiv:2108.09484.

40. Brama, E.; Peddie, C.; Wilkes, G.; Gu, Y.; Collinson, L.; Jones, M. ultraLM and miniLM: Locator tools for smart tracking of fluorescent cells in correlative light and electron microscopy. *Wellcome Open Res.* **2016**, *1*, 26. [CrossRef] [PubMed]

41. Guskin, S.; Wasserblat, M.; Wang, C.; Shen, H. QuaLA-MiniLM: A Quantized Length Adaptive MiniLM. *arXiv* **2023**, arXiv:2210.17114.

42. Vergou, E.; Pagouni, I.; Nanos, M.; Kermanidis, K.L. Readability Classification with Wikipedia Data and All-MiniLM Embeddings. In Proceedings of the Artificial Intelligence Applications and Innovations, León, Spain, 14–17 June 2023; pp. 369–380.

43. Chakravarthi, B.R.; Jagadeeshan, M.B.; Palanikumar, V.; Priyadharshini, R. Offensive language identification in dravidian languages using MPNet and CNN. *Int. J. Inf. Manag. Data Insights* **2023**, *3*, 100151. [CrossRef]

44. Quyen, V.T.; Kim, M.Y. MPNet: Multiscale predictions based on feature pyramid network for semantic segmentation. In Proceedings of the 2023 Fourteenth International Conference on Ubiquitous and Future Networks (ICUFN), Paris, France, 4–7 July 2023; pp. 114–119. [CrossRef]

45. Qianmin, S.; Wei, P.; Xiaoqiong, C.; Hongxing, L.; Jihan, H. COVID-19 clinical medical relationship extraction based on MPNet. *IET Cyber-Phys. Syst. Theory Appl.* **2023**, *8*, 119–129. [CrossRef]

46. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv* **2020**, arXiv:1910.01108.

47. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. The DistilBERT Model: A Promising Approach to Improve Machine Reading Comprehension Models. *Int. J. Recent Innov. Trends Comput. Commun.* **2023**, *11*, 293–309. [CrossRef]