National College of Ireland

Comparative Evaluation of Resume-Job Matching Using Transformer-Based Semantic Similarity and TF-IDF

MSc Research Project

Master of Science in AI

Alex Power

24100421

School of Computing

National College of Ireland

Supervisor: SM Raza Abidi

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Alex Power

Student ID: X24100421

Programme: MSc in Artificial Intelligence

Module: Practicum

Supervisor: SM Raza Abidi

Submission Due Data: December 12 2025

Project Title: Comparative Evaluation of Resume–Job Matching Using Transformer-Based Semantic Similarity and TF-IDF

Word Count: 6601 **Page Count: 19**

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Alex Power

Date: 10/12/2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	✓
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	✓
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	✓

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Comparative Evaluation of Resume–Job Matching Using Transformer-Based Semantic Similarity and TF-IDF

Alex Power

24100421

Abstract

This research proposes a candidate-centric system for matching resumes with relevant job descriptions using transformer-based sentence embeddings. Using semantic similarity scoring, this research removes the need for training data and improves relevance in job recommendations. The proposed system uses Sentence-BERT to embed both the given resume and a subset of job descriptions, comparing them using cosine similarity. This will be done using TF-IDF as a baseline for comparison. They will be evaluated using a combination of quantitative and qualitative metrics. Quantitative metrics include nDCG, MRR, Precision@K, Recall@K, Spearman Correlation, Runtime and Memory Usage. The qualitative evaluation consists of feedback based on participant perception of ranked job descriptions gathered from individuals with experience in hiring, HR or related experience. Across all resume domains, SBERT consistently outperformed TF-IDF on ranking accuracy, achieving higher MRR, nDCG, and stronger alignment with human-judged relevance. SBERT also received higher average participant relevance scores in four of five domains. Although TF-IDF demonstrated significantly faster runtime, it required considerably more memory, whereas SBERT exhibited slower inference but lower and more stable memory usage. The findings suggest that transformer-based models provide more accurate and semantically meaningful job recommendations without domain-specific fine-tuning, highlighting their suitability for recruitment contexts where contextual understanding is essential.

Keywords: Resume-Job Matching, Transformer-based Models, Semantic Similarity, SBERT, TF-IDF

1 Introduction

Artificial intelligence (AI) and Natural Language Processing (NLP) technologies are increasingly being used to automate candidate screening and job matching (*Khan et al. 2025)*. These screening tools can parse and classify candidate resumes at a rate employees cannot,

streamlining this process. While this increase in efficiency is a primary factor leading to the increase in utilization of these tools, cost-related factors are another important consideration. Utilization of these tools can allow for employees in recruitment or human resources to significantly reduce time spent sifting through the first round- of applicants. Recent surveys show an increase of use in these tools within recruitment departments, with a LinkedIn survey from 2024 indicating that 27% of large organisations now use a form of automated recruitment technology, with six out of ten organisations surveyed "optimistic about AI in recruitment" (*Linkedin Talent Solutions 2024)*.

Many modern AI screening tools parse resumes into structured representations and classifies applicants based on a predefined criterion *(Elsafty et al. 2018).* Traditional systems used for this task would rely on keyword matching, or machine learning models trained on labelled data to indicate whether candidates were applicable to the given role (*Qaiser and Ali 2018).* These traditional models often use features from the job description such as skills, and years of experience against resumes to classify. While the utilization of NLP in these modern tools has shown an increase, limitations still exist surrounding unstructured input data and supervised learning approaches (*Sasi 2024).*

The key issue in creating these tools arises in the limited access to labelled recruitment data, from regulations such as GDPR and ethical concerns surrounding privacy (*Sartor and Lagioia 2020).* This results in some modern systems being trained on undiversified data, leading to potential bias in the system. This can lead to applicants being misclassified, leading to significant ethical issues (*Mujtaba and Mahapatra 2019).* Further issues arise with these systems when applicants have non-standardised resumes or are applying for a role in a niche career with a small subset of the training data if any at all. Contrasting this, applicants applying for a role in an industry they have no professional experience in may be unfairly classified as inadequate even when having the necessary skills and education, due to poor feature extraction prioritising past work experience over skills and education (*Rigotti and Fosch Villaronga 2024).* These challenges highlight the limitations of current AI-based screening systems, highlighting the importance of further research on applying different types of models on the task of resume-job description matching.

The motivation for this work comes from the recent well-documented advancements in semantic similarity modelling using transformer architectures such as SBERT (*Reimers and Gurevych 2019).* These transformer models have been widely adopted in document retrieval tasks and sentence similarity, without extensive research conducted into candidate-job matching possibilities (*Gillioz et al. 2020).* These models have been shown effective at semantic search and sentence similarity, allowing for dense embedding and ranking systems far beyond the simple word matching of traditional models (*SBERT 2025).* Previous research has shown an increase in performance when compared to traditional models in domain-specific retrieval, with NASA design information retrieval being a primary example (*Walsh and Andrade 2022).* While research has been conducted comparing these two models, further research is required in more varied domains such as resume-job descriptions matching where bias and mistakes can have significant legal and ethical implications. This research aims to

provide further understanding to how these transformer models can perform resume-job matching while compared to a TF-IDF model.

The purpose of this research is to flip the perspective of job matching. Rather than screening candidates for a specific job, the proposed system will evaluate how the given resume semantically aligns with a wide range of job postings. This research will document two approaches, a transformer-based model and a traditional statistical model based on term frequency-inverse document frequency (TF-IDF). TF- IDF is a well-established technique in information retrieval, representing documents as sparse vectors weighted by word importance (*Qaiser and Ali 2018).* The key limitation of this method is its inability to capture deep semantic meaning. Sentence-BERT (SBERT) is a transformer-based language model designed for creating high quality sentence embeddings to capture contextual meaning in natural language (*Reimers and Gurevych 2019).* The system for this research will be created using SBERT as the transformer-based model in contrast to the TF-IDF model.

Research Question:

How effectively can transformer-based embedding models match candidate resumes to job descriptions using semantic similarity scoring, and how do these models compare to traditional techniques in terms of accuracy and interpretability?

The proposed system involves creating a prototype that ingests a resume, embeds it using a transformer- based model and computes similarity scores against the dataset of job-posting descriptions. The system will return the ranked list of matching jobs. The system will be evaluated using quantitative metrics to assess the performance, with a qualitative evaluation consisting of a human reviewing aspect based on the perceived match relevance.

The rest of the article is arranged as follows:

In Section **2**, an overview of the current literature will be extensively discussed, highlighting some of the most relevant papers to this topic. This section will explain their capabilities, limitations, and possible improvements. Section **3** outlines the methodology used to carry out this research. This section will detail the research approach by outlining the chosen datasets, system architecture, embedding models, tools, frameworks and resources. There will be both qualitative and quantitative evaluation metrics used to assess the performance of the system. Due to the qualitative evaluation requiring human participants, section 3 will discuss the ethical considerations of this research. Section **4** outlines the evaluations carried out as part of this research. This section will detail the qualitative and quantitative evaluation results of both models. A thorough comparison will discuss the benefits and drawbacks of each model and how they performed at matching resumes to job descriptions. Section **5** will conclude this report by discussing the outcomes of the research, and possible future work. While discussing the outcomes of this research, relevant limitations will be noted and expanded on to inform future researchers of possible aspects to consider or focus on.

2 Related Work

A. Early Approaches to Automated Resume-Job Matching

Initial approaches to automated recruitment tools often relied on lexical matching. Candidates were scored based on literal word overlap with systems built on TF-IDF or keyword-based retrieval which treated resumes and job descriptions as collections of words. The key limitation of these systems was how two candidates could be scored very differently depending on their choice of words in their resumes. While these systems didn't have much computational requirements, this limitation could cause a variety of issues. Research have discussed how these models can overlook relationships between words that have the same meaning, that aren't spelled the same way, leading to candidate rankings suffering when compared to human judgement (*Khan et al., 2025)*.

The introduction of dense vector representation (Word2Vec and Doc2Vec) to encode documents in continuous semantic spaces, highlighted a potential method of alleviating these limitations. Early insight into the value of semantic information came from *Elsafty et al. (2018)* when they highlighted the improvements in matching accuracy after embedding jobs and resumes, when compared to keyword retrieval. Further research discusses the limitations of these static embeddings, as they are fixed representations, contextual nuance cannot be interpreted (*Nigam et al., 2019)*. This limitation is heightened in relation to long job postings or resumes. With this in consideration, while static-embedding systems are an improvement over direct lexical matching, they are still unable to capture any meaning behind qualifications and experience.

B. Contextual Transformer Models in Recruitment

Transformer architectures like BERT was a significant improvement in modelling the meaning of terms in candidate resumes. Transformer are capable of considering the context of each word, allowing these models to understand different phrases can express equivalent skills, experience or education. Research applying BERT to recruitment tasks has shown results more aligned with human judgement, signifying further improvements over past methods (*Devlin et al., 2019, Khan et al., 2025*).

A study from *Bhatia et al. (2019)* showed BERT's capability at directly estimating the compatibility between a resume and job by treating the task as a paired sequence classification problem. This work has shown notable improvements in automating initial screening in recruitment, highlighting the potential of transformer-based models. Recent studies such as *Jirjees et al. (2025),* have built upon this research, highlighting the advantages of contextual embeddings by showing how they can outperform classical features in classification and ranking, even in situations where candidates phrasing isn't identical to job description terminology.

C. Domain-Adapted Transformer Models

The growth of transformers in natural language processing has led researchers to begin adapting them for specific tasks, such as recruitment. One noteworthy model is conSultantBERT, which was trained on hundreds of thousands of resume-vacancy pairs to learn highly specialised representations for matching *(Lavi et al., 2021)*. The "Siamese" architecture enables retrieval that significantly outperforms TF-IDF baselines.

Additional work has incorporated "structured labour-market knowledge" into transformer pipelines. *Rosenberger et al. (2025)* developed CareerBERT, which aligns resumes and job descriptions using a shared embedding space, derived from the ESCO taxonomy of skills and occupations. By defining skills with standardised descriptions, the model was capable of strong agreement with human evaluators, addressing further limitations from traditional methods.

Further research from *Bevara et al. (2025)* proposes the Resume2Vec framework. This framework integrates multiple transformer encoders, allowing the generation of resume embeddings for large scale systems. The results further suggested a clear performance gap between the legacy methods and transformer-based retrieval, with improvements noted across diverse job domains.

D. Contrastive Learning and Dense Retrieval Techniques

Recent research explores contrastive learning, which trains models to distinguish between relevant and irrelevant resume-job pairs using both positive and negative examples. Research from *Yu et al. (2024)* proposed ConFit, a dual-encoder framework that applies contrastive objectives to produce more discriminative representations for retrieval. This approach incorporates synthetic data produced by LLMs to combat the scarcity of labelled recruitment data. ConFit has shown strong performance in ranking accuracy, demonstrating the practicality of transformer-based dense retrieval through vector search libraries.

E. Findings and Research Gaps

Current research consistently highlights semantic, transformer-based representations over traditional TF-IDF methods for resume-job matching. These contextual models are more capable and consistent at capturing the skills, experience and qualifications of resumes, allowing for more accurate rankings that reflect human judgement.

However, there are still a number of gaps in research. One key gap is the dependence on large training data, with most capable models being trained on private datasets which are unavailable to smaller organisations. Further research should be conducted on the performance of these transformer-based models on public datasets. Additional gaps surrounding transparency and interpretability need to be explored. Decisions in recruitment require justification and need to remain unbiased. Transformer-based systems should explore explainability features to ensure bias is not present, with fairness and regulatory compliance ensured.

Given these gaps in research, this study aims to contribute by exploring how well a general-purpose transformer embedding model can perform in comparison to a TF-IDF baseline, in a smaller-data context. This study will analyse whether semantic similarity improves ranking

quality without domain-specific fine-tuning. This comparison aims to provide insights into the practicality of deploying embedding-based matching methods in recruitment situations, where a large amount of data is not readily available.

3 Methodology

A. Research design

When conducting this research, a Design Science Research (DSR) methodology was closely followed *(Peffers et al., 2007)*. This methodology focuses on developing, implementing and evaluating a method of addressing a defined research problem. Within this study, the method for addressing the research problem is the system designed to compare the effectiveness of transformer-based semantic embeddings and traditional statistical approaches for resume-job description matching. The chosen methodology combines both experimental and comparative aspects, with the primary objective to evaluate the performance and interpretability of the two models when tasked with the same resume on the same dataset.

The chosen methodology was outlined shown in **Figure 1**. The first stage included identifying the problem and outlining the objective of the research. The current limitations of similar systems were noted, with the aim and evaluation goals of the created system defined. The second stage included the design and development of the system. This stage was more code-focused, with the models being set up, integrated into one pipeline, and ran from a Streamlit dashboard for simplicity when it came to the human participants. The third stage included running the evaluations and analysing the results. Quantitative results include MRR, nDCG, Precision@K, Recall@1/3/5, Spearman Correlation, Runtime, and Memory usage. The qualitative evaluations included hosting a small group of five participants to evaluate the system using a Likert Scale of 1-5. The final stage included discussing the results, the research and the future work of AI in resume-job matching.

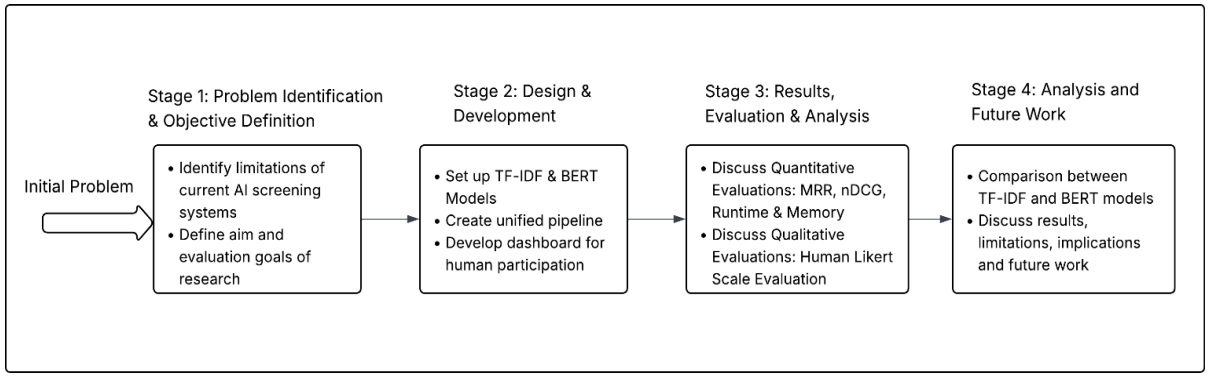

Figure 1: Methodology Process

The system with both models was created using Python. While a full list of used libraries will be noted in the Read.ME section of the coding artefact, libraries include *pandas, numpy, scikit-learn and sentence-transformers*. Both models were evaluated using the same job description dataset and resume for each comparison to ensure fairness and accuracy.

B. Datasets & Preprocessing

The chosen datasets for this research are the "Kaggle Resume Dataset", and the "LinkedIn Job Postings Dataset" *(Koneru and Yu Zou 2025) (Bhawal, 2025).* Both datasets were cleaned and anonymised before use to ensure ethical compliance and data-protection standards. In addition to this, a handful of custom-made resumes were made for evaluation and further testing for real-world scenarios.

The preprocessing steps for the datasets included the removal of null rows, and the removal of HTML tags that were left in the dataset. The final dataset was used for evaluations, with one resume being matched with all job descriptions in the dataset.

C. Evaluation Metrics

This research includes both quantitative and qualitative metrics to provide a comprehensive assessment of the two models.

The quantitative metrics will allow for the models to be evaluated using standard information-retrieval metrics. Runtime and memory metrics will be used to assess the resources used by each model, offering a more comprehensive view of each model. Mean Reciprocal Rank (MRR) will be used to evaluate the ranking accuracy based on the most relevant results. Normalised Discounted Cumulative Gain (nDCG) will be used to account for the ranking and positioning of job descriptions based on their relevance. Precision@K will be used to evaluate the accuracy of relevant matched jobs. Recall@1/3/5 will be used to evaluate the position of ranked jobs. Spearman correlation will be used to rank the jobs in comparison to how the human participants ranked matched jobs, highlighting the similarity each model has with human perception.

The qualitative evaluation aspect of this research will come from a group of participants ranking the results from both models. The participants will assess the ranked outputs on a dashboard, ranking how relevant they are on a Likert scale ranging from 1-5, with 1 being "Irrelevant", and 5 being "Highly relevant". This evaluation aims to capture the subjective perceptions of semantic alignment which would be more difficult to gather through quantitative evaluations. It will provide an insight into how this may work in a real-world context.

D. Ethical Considerations

The datasets used in this research are publicly available on Kaggle and have been anonymised. This ensures no personal or private information is processed or used for this research. The qualitative evaluation involved voluntary peer participation, with all feedback being collected anonymously without any personally identifiable data. The qualitative feedback was collected only after receiving the participant's informed consent. This research is in compliance with ethical guidelines and adheres to GDPR standards set regarding using data. This research aims to be transparent, reproducible and responsibly use AI.

4 Implementation & System Design

A. Implementation Method

Two models were chosen for comparison in this research, a transformer-based model (SBERT), and a traditional statistical model (TF-IDF).

The specific sentence-BERT (SBERT) model chosen was the *MiniLM-L6-v2*. This model was used to generate dense vector embeddings for resumes and job descriptions. Cosine similarity was used to compute semantic similarity between the chosen resume and all available job descriptions.

The TF-IDF model was used as a baseline comparison for SBERT. The TF-IDF approach allows text to be represented as sparse weighted vectors. Similarity between resumes and job descriptions were computed again using cosine similarity. An example of this would be "Python, Excel, Tableau" in roles such as finance or programming. Both models were integrated into the same evaluation pipeline to ensure a direct comparison under the same conditions.

B. System Design

The system was designed with non-technical users in mind. An easy-to-understand dashboard was incorporated in the system to allow the ranking to be done within the system, with the results saved to an excel file. This allows the participant to directly save their results for each job description, for a given resume. Additionally, the dashboard allows for a resume to be uploaded, with a button below to run the models.

Figure 2 depicts an image of the qualitative evaluation dashboard. This dashboard was created for ease-of-access to the participants of this research. This dashboard allows for the running of both models for one resume against the job descriptions dataset. In addition to this, on the sidebar, the user is able to load precomputed resumes allowing for the qualitative research to be carried out, without requiring the models to be ran in between each aspect of the research. For this research, the results from five resumes are precomputed for each model, with a total of 50 jobs ranked by relevance.

Figure 3 depicts the dashboard once the precomputed results are shown, the user is able to click into each ranked job, read the description and vote using the radio buttons whether the job is 1: Irrelevant, 2: Somewhat Relative, all the way to 5: Perfect Job for the given Resume. Additionally, there is a space for a comment if the participant's feel they need to make a comment. An analysis of the comments would have been placed in the analysis aspect of this research report, had any participants noted any. Although, none of the participants wrote a comment on any job descriptions, this aspect still allowed the participant more choice with their answers, potentially leading to a more informed research study.

For this research, the system design was broken down into four processes. The first stage included clearly identifying the problem and outlining the objectives. The second stage included designing and developing the system. The third stage included evaluating the system and assessing its performance. The fourth stage included analysing the outcomes of both evaluation types and discussing the implications they have on future research.

Figure 2: Image of Qualitative Evaluation Dashboard

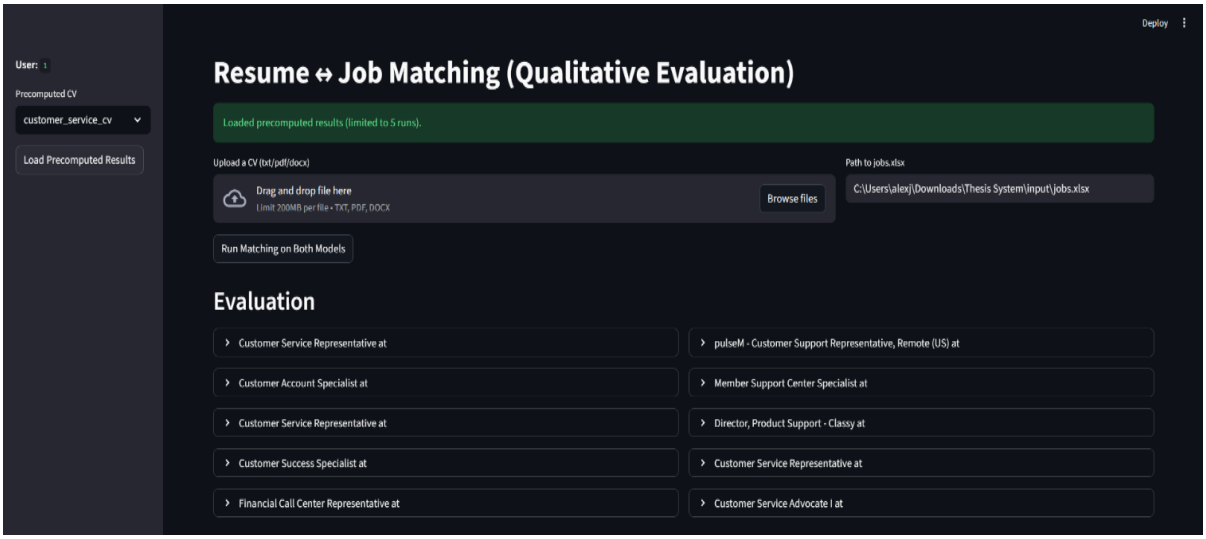

Figure 3: Precomputed Results Displayed on Dashboard

5 Results, Evaluations & Analysis

This section covers the results of the resume-job matching system's evaluations, across quantitative, qualitative, and technical metrics. The aim of these evaluations is to assess the performance of the system, highlighting how effect the TF-IDF and SBERT based models rank relevant job postings per resume. The qualitative evaluation consists of the participant feedback on model output relevance, providing insight into a practical use of the models, past metrics. Technical performance includes the specifications of the computer ran on with the runtime, and memory consumption, to further understand the use of these models in a real-life application. This section ends with a comparison between the two models, highlighting the potential implications each model has for resume-job matching applications.

5.1 Quantitative Evaluation

A. Ranking Accuracy Metrics

The effectiveness of both models was assessed using a variety of metrics, with a summary of these metrics being displayed above in table one. These metrics include Mean Reciprocal Rank (MRR), Normalised Discounted Cumulative Gain (nDCG), Precision@K, Recall@K, and Spearman rank correlation. Initial findings found skewed metrics when based without human participant feedback as a baseline. The following results are based on the relevance scores from the participant assessment aspect of this research to ensure a realistic measure of performance shown in **Table 1**.

As shown in Table 1, across the five primary resumes used for this research, SBERT consistently achieved stronger ranking results than the TF-IDF model. SBERT achieved an MRR of 1.0 for three of five resumes, which indicates that the highest rated job for that resume appeared first. TF-IDF only achieved an MRR of 1.0 for one resume, which SBERT also achieved 1.0. The other MRR values for the TF-IDF model were 0.33333 for Customer Service, 0.5 for Executive Assistant, 0.5 for Legal, and 0 for English Teacher. This highlights the difference between the two models at ranking the most relevant job.

Table 1: Summary of Quantitative Metrics

CV	Model	MRR	nDCG@10	Precision @1	Precision @3	Precision @5	Recall @1	Recall @3	Recall @5	Spearman
Customer Service	SBERT	1.0000	1.0000	1.0	1.0000	1.0	0.2000	0.6000	1.0	0.7071
Customer Service	TFIDF	0.3333	0.9026	1.0	0.6667	0.8	0.2500	0.5000	1.0	-0.5798
English Teacher	SBERT	1.0000	0.9691	1.0	0.6667	0.6	0.3333	0.6667	1.0	0.6325
English Teacher	TFIDF	0.0000	0.9843	1.0	0.6667	0.4	0.5000	1.0000	1.0	0.7906
Executive Assistant	SBERT	0.3333	0.9478	1.0	1.0000	1.0	0.2000	0.6000	1.0	-0.2887
Executive Assistant	TFIDF	0.5000	0.9059	0.0	0.6667	0.8	0.0000	0.5000	1.0	-0.5270
Graphic Designer	SBERT	1.0000	0.9902	1.0	1.0000	1.0	0.2000	0.6000	1.0	0.5774
Graphic Designer	TFIDF	1.0000	0.9818	1.0	1.0000	1.0	0.2000	0.6000	1.0	0.0000
Legal	SBERT	0.3333	0.9582	1.0	1.0000	1.0	0.2000	0.6000	1.0	-0.1814
Legal	TFIDF	0.5000	0.9703	1.0	1.0000	0.8	0.2500	0.7500	1.0	0.6708

Figure 4 shows the nDCG@10 results follow this pattern. SBERT was capable of consistently ranking relevant job descriptions with the lowest being 0.9478 and the highest being 1.0. The results from the TF-IDF model ranged from 0.9026 to 0.9843, further underperforming in comparison to SBERT. While the SBERT model yielded better results, the results from the TF-IDF still closely align and show promise.

Figure 5 Precision and Recall were used at a 1, 3, and 5 to give a clear overview of performance. Precision@1 saw SBERT achieving 1.0 for all five resumes, while TF-IDF achieved a result of 1.0 in four resumes. Precision@3 and Precision@5 followed the same trend, with SBERT achieving higher results. As seen in figure six, Precision@5, shows SBERT's results higher in four categories. Recall@5 showed both models achieving 1.0 for all resumes, likely from the resumes being tested already containing relevance. In contrast,

Recall@3 and Recal@1, SBERT achieved more consistent early retrieval of more relevant job matches.

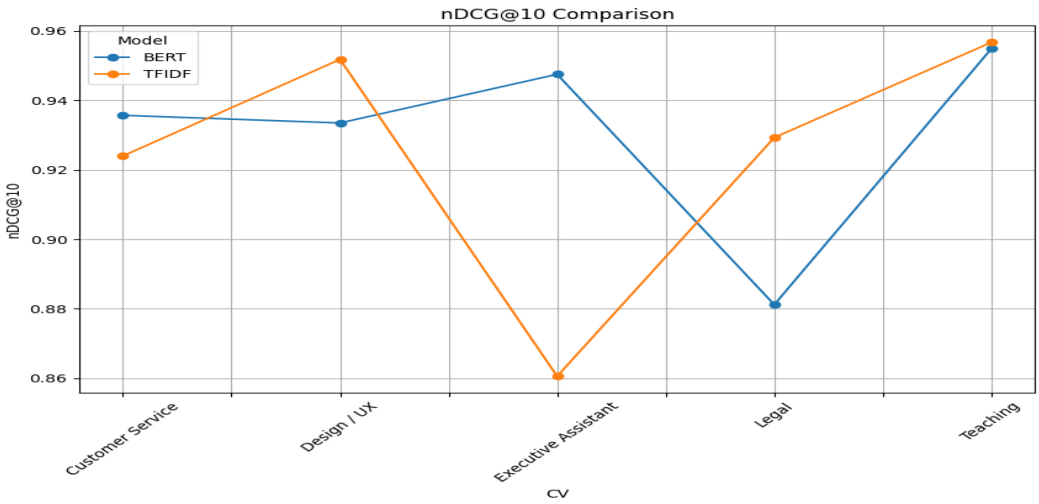

Figure 4: Graph of nDCG@10 comparison between models

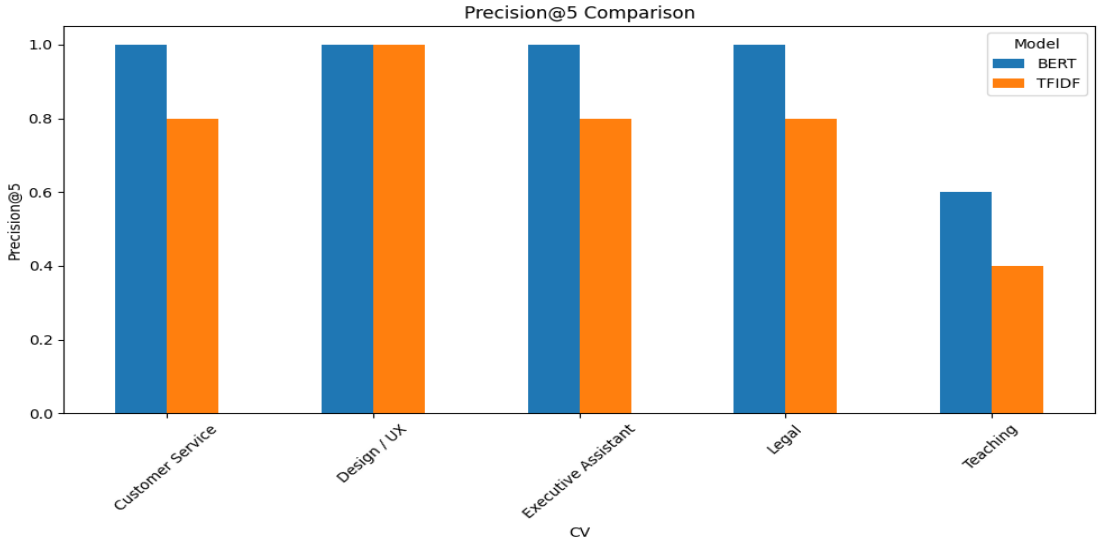

Figure 5: Graph of Precision@5 Comparison

Figure 6 shows the Spearman rank correlation shows the model's ordering compared with human choices. SBERT achieved similar rankings compared to human ranking with four out of five resumes aligning, while TF-IDF demonstrated near-zero results in several cases. This highlights SBERT's ability to choose jobs for a resume more comparable to the choice a human would.

These metrics indicate that the SBERT based model consistently provides more accurate job rankings across various resume types. This section highlights SBERT as the better performing model at the task of ranking job matchings for a given resume within this study.

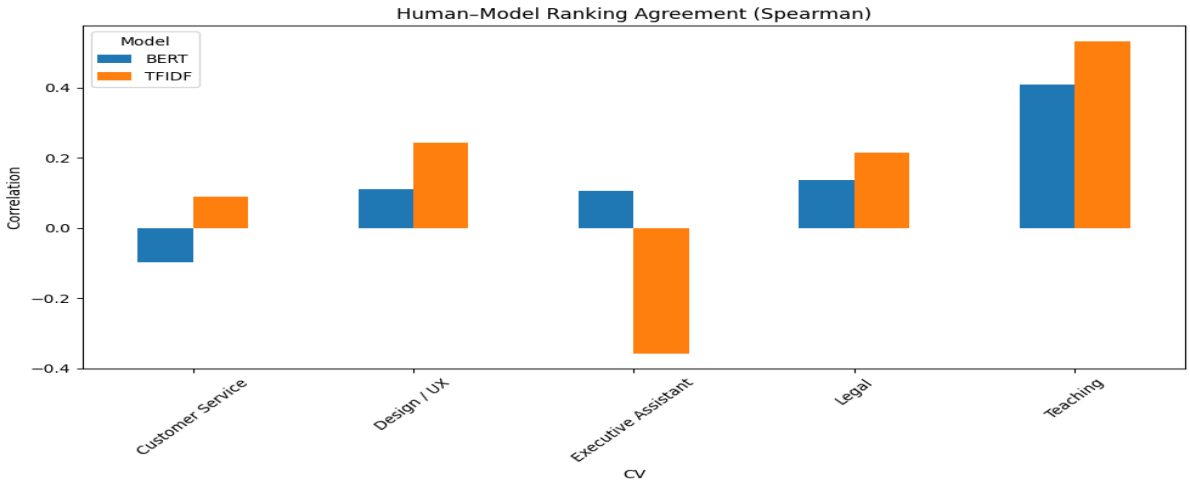

Figure 6: Spearman Correlation

B. Model Performance

The performance of each model varied across the different domains of resume, showing weaker results in more diverse categories. SBERT showed consistent and reliable performance across all resumes, accurately ranking relevant jobs high even with complex and diverse wording compared to the resume. This consistency highlights SBERT's ability to understand semantic similarity beyond simple *text*-matching. In contrast, the TF-IDF showed unreliable results with more sporadic behaviour. TF-IDF performed well with resumes where it was capable of text-matching such as *Graphic Design* and *English Teacher*, but performance dropped significantly for resumes with broader terms and more diverse language such as *Customer Service*.

Another key observation is model performance in generalisation of jobs. SBERT was capable of ranking jobs based on relevance in more reliably than the TF-IDF model. TF-IDF often had a drop in relevance after the first few matches leading to weaker results following the first few. This highlights the TF-IDF's inability to understand broader skills and role similarity beyond text-matching.

The metrics and analysis above clearly outline SBERT as the higher performing model, with consistence and reliability across a variety of diversified resumes. While TF-IDF under-performed in comparison, it is important to note that it performed well in less diverse domains, where text-matching was available.

5.2 Qualitative Evaluation

A. Human Participant Results

The qualitative evaluation consisted of five participants ranking the output from both the TF-IDF and SBERT models across five resumes. Each participant reviewed five job recommendations from each model, for each resume, resulting in a total of fifty rankings per participant. The participants' relevance ranking ranged from 1 (irrelevant) to 5 (perfect job), allowing each participant to rank under the same rules. This aspect of the research aimed to understand how closely each model's ranking aligned with human judgement. The participants

range from a variety of backgrounds relating to hiring. This enabled them to make their decisions with the informed context of what to look for when hiring a candidate based on the resume.

As shown in **Figure 7**, participants consistently ranked the SBERT results as more relevant compared to the TF-IDF model. SBERT produced a higher number of rankings in the 4 and 5 across all resumes. SBERT can be seen to outperform TF-IDF in four categories, often by a significant amount. Participant three is the only participant who found the TF-IDF model provided higher relevance outputs. This displays the subjectivity of the qualitative evaluation. In contrast, Participant four had the largest difference between models, with SBERT results being ranked much higher than TF-IDF. This indicates the ranked jobs from SBERT were more aligned with participants expectations for a given resume. In contrast, TF-IDF only demonstrated consistency in a few domains, specifically the *Graphic Designer* and *English Teacher*. This is likely from the keyword overlap found in these domains, resulting in more matches and better ranking. In domains where there is more contextual understanding required such as *Customer Service* or *Executive Assistant*, TF-IDF had less consistent results. This pattern aligns with the quantitative evaluations, displaying TF-IDF's inability to capture deeper meaning as its main limitation. These findings suggest SBERT's ability to understand semantic similarity with deep contextual meaning, results in much better results.

Participants' rankings for SBERT outputs were much more consistent than TF-IDF, with very few results being one or two. In contrast, the results from TF-IDF were less consistent, varying for all resumes. This likely indicates TF-IDF was suggesting weaker jobs.

Figure 8 shows the average human rating per resume domain for each model. The results show SBERT achieving higher than TF-IDF in four categories, following the previous trend. These results reflect SBERT's ability to interpret contextual meaning within the resumes and match to jobs without the requirement of explicit keywords. The *Graphic Designer* resume, titled above as *Design / UX* shows identical ratings. This is likely from the extensive amount of keyword overlap in this domain, including tools, software and standardised vocabulary. The *English Teaching* resume shows the largest gap between models, likely arising from the fact that TF-IDF often matched teaching jobs for other subjects with the *English Teacher* resume. TF-IDF used keyword overlap to primarily match this resume with any job that had teaching, with one of them being a science teacher. This graph further reinforces SBERT's ability to perform consistently across a variety of domains, while highlighting TF-IDF's ability to only excel in very standardised domains.

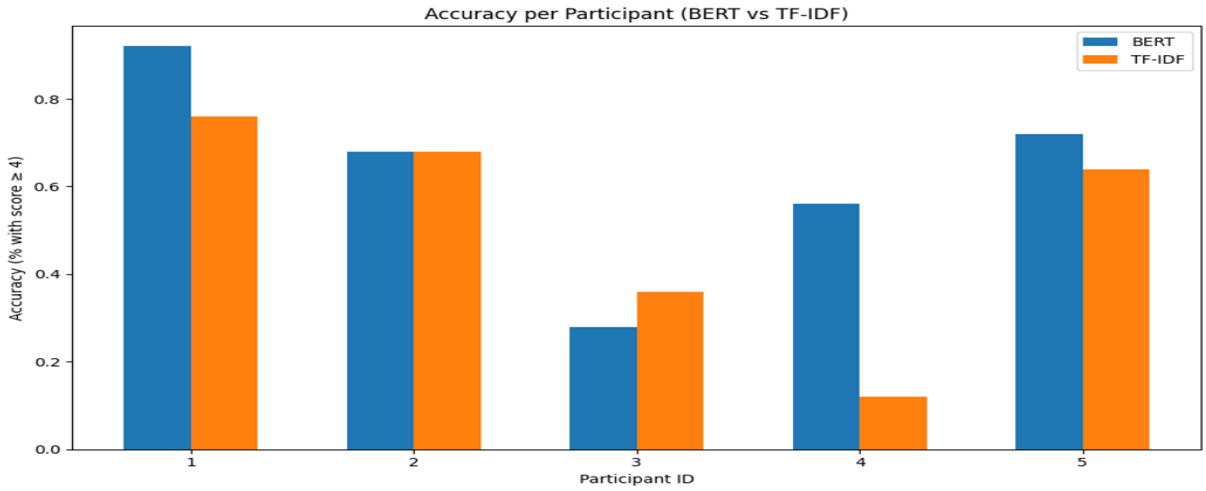

Figure 7: Graph of Accuracy per Participant per Model

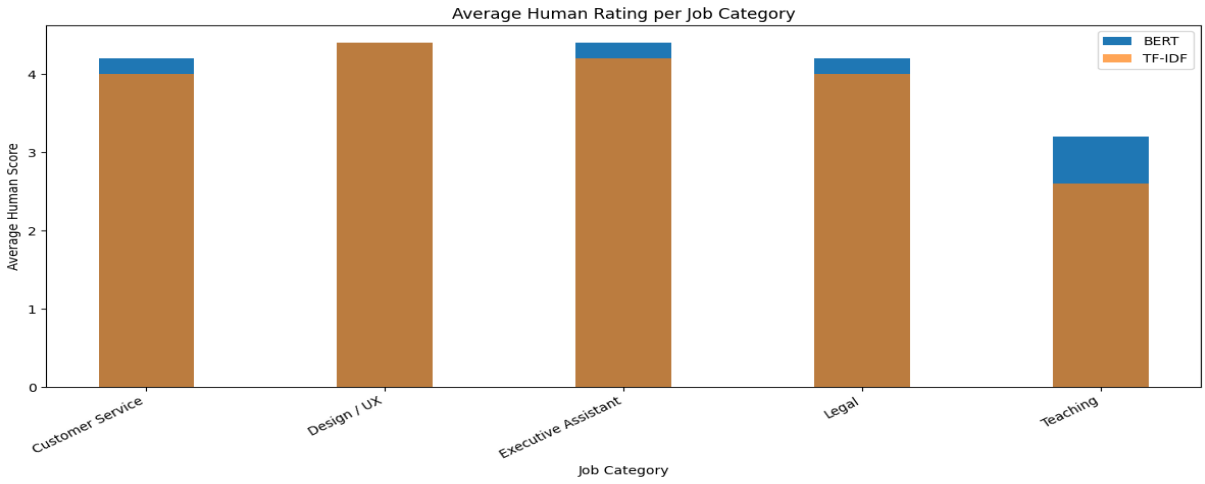

Figure 8: Graph of Average Human Rating per Resume

B. Example Result of Good Ranking

SBERT's results from the *Customer Service* resume were often ranked with scores of four or five. These rankings include roles such as *Customer Support Representative*, *Member Support Specialist*, and *Customer Service Advocate*. These results were highly relevant to the resume, in the eyes of the participants and the decision from SBERT. Additionally, the ranking in which they were placed by SBERT, closely followed the rankings of the participants, with the top recommendation being ranked five. The combination of high human relevance scores and consistent ranking position shows SBERT's ability to prioritize jobs that best match the resume, even when job titles differ.

While TF-IDF was outperformed by SBERT, the model displayed good results in certain domains. In the results from *Graphic Designer* resume, the participants ranked quite high scores of four to five in a number of results. Jobs in this domain would have a lot of keywords overlap from required software or type of work, leading to improved results. This example displays how TF-IDF can be effective in certain domains, even if SBERT is capable of being effective in more.

C. Error Analysis

A review of participant ratings across all resumes and models found some inconsistencies between participants. This was most prevalent in mid-relevance roles where one participant ranked two and another ranked three, or one ranking three and another ranking four. These differences are expected as not every participant will rate a resume the exact same way. This qualitative evaluation inherently carries subjectivity, leading to the choice of a group of five participants rather than one.

5.3 Technical Performance

Both models were evaluated using runtime and memory usage during the ranking of job descriptions for each resume. These results allow for an insight into the computational requirements of each model, offering a more comprehensive understanding of the efforts when deploying, scaling and managing each model. The following technical evaluations were run on a computer with the following specifications. Graphics Card: NVIDIA GeForceRTX 3070 8GB RAM, Installed Ram: 48GB, Processor: Intel® Core™ i7-10700 CPU @ 2.90GHz.

A. Runtime Performance

As shown in **Table 2**, TF-IDF consistently outperformed SBERT for runtime. TF-IDF completed each ranking in under a minute (ranging from 54 - 65 seconds), while SBERT took around six minutes on average (ranging from 283 - 381 seconds). The difference highlights the further computational requirements for the SBERT model. It is important to consider that the times may fluctuate dependant on the hardware the models are ran on, with more powerful hardware, both models could run faster.

Table 2: Runtime Results

	Model	
CV	**SBERT**	**TFIDF**
Accounting	381.0412	60.7978
Customer Service	325.4844	58.0783
English Teacher	328.3806	65.6457
Graphic Designer	336.4390	59.3214
Lawyer	283.9983	54.1591

B. Memory Usage

Table 3 shows the memory usage of the models showed opposing results to runtime, with TF-IDF consistently requiring more memory than SBERT. The memory usage of TF-IDF ranged from 565 MB to 705 MB dependant on the resume, while SBERT's was more consistent and ranged from 278 MB to 309 MB. The higher memory usage from storing the weights for each term for all job descriptions. With a larger dataset, the memory usage is likely to increase significantly in comparison to SBERT. In contrast, SBERT loads a fixed size model into

memory, processing text into embeddings without requiring a large amount of space, leading to a more efficient method of memory usage.

Table 3: Memory Usage Results

	Model	
CV	**SBERT**	**TFIDF**
Accounting	303.4547	571.3190
Customer Service	278.2644	566.0340
English Teacher	278.2648	565.9234
Graphic Designer	278.2645	565.9451
Lawyer	278.3547	566.2878

C. Technical Considerations

Each model outperforms the other for each technical evaluation carried out. While TF-IDF ran significantly faster than the SBERT model, it required much more memory. The key limitation of the TF-IDF model is the high memory usage which will grow on large datasets. The key limitation for SBERT is its slow inference time which could lead to issues dependant on the number of resumes or job descriptions required for matching. In terms of technical requirements, each model has its own use-case and should be chosen for specific situations.

5.4 Comparative Analysis

A. Model Comparison

When comparing the results from both models, it's clear that SBERT has an advantage over the TF-IDF model. In terms of ranking quality, placement and in alignment with human judgement, SBERT produced more consistent and reliable results. SBERT's results showed higher MRR, nDCG, and a positive Spearman correlation for most resumes, highlighting its performance in order ranked. While TF-IDF performed well in domains with high lexical-overlap where text-matching was possible, it was significantly less reliable in roles requiring broader semantic understanding. SBERT showed an understanding in all resumes, with the ability to understand the broader context for more complex resumes such as Customer Service, where TF-IDF struggled.

B. Implications

The strengths and weaknesses of both models have practical implications in real-world applications. In hiring workflows where ranking accuracy and semantic similarity are crucial, SBERT is the more suitable model, based on most metrics. While TF-IDF can be used for completing the task faster, SBERT is a more efficient, more accurate and more reliable method to parse resumes and job descriptions. A possible use for both models could be a hybrid strategy, using SBERT for the initial parsing, following up with the TF-IDF model for efficiency on the final resumes. Although further research is required for hybrid approaches and their success.

5.5 Summary of Findings

Across all evaluations conducted in this study, the SBERT transformer-based model consistently outperformed the traditional TF-IDF approach for resume-job matching. Quantitative metrics highlighted SBERT's ability to rank relevant job roles using semantic relevance with closer alignment to human-judged relevance than TF-IDF. Qualitative feedback from participants supported these findings with four out of five results from participants suggesting SBERT was recommending more relevant roles. Technical evaluations highlighted the trade-offs between the two models with TF-IDF running a resume against the job description dataset at a much faster rate, with higher memory usage. In contrast, while SBERT took much longer to run, the memory usage was lower and more consistent across resumes. Overall, the findings from the evaluation indicate that SBERT is more suitable at the task of resume-job matching tasks, specifically in domains that require more contextual understanding for ranking, and where accuracy is crucial.

6 Conclusion and Future Work

6.1 Study Conclusions

This research report aimed to evaluate to what extent, a transformer-based semantic similarity model, SBERT (MiniLM-L6v2), can match resumes to job descriptions in comparison to a traditional TF-IDF model. This research used quantitative, qualitative and technical evaluations to offer a comprehensive explanation of how each model performs when tasked with matching jobs to a given resume. The results show SBERT consistently outperforming the TF-IDF model, highlighting its ability to use transformer-based semantic embeddings, over traditional keyword overlap.

The system created for this study contributes to existing research, allowing for the testing of both models in future qualitative studies from an easy-to-understand dashboard, with the ability to preload resume results, rank relevance of jobs and save them for analysis. The qualitative evaluations from this research have shown that even without domain-specific fine-tuning, the SBERT model is more capable of producing results aligned with human judgement. This further highlights the strength of the semantic embeddings in capturing both explicit and implied skills, a feature the traditional model could not show.

Additionally, this research aimed to discuss important considerations when choosing how to create systems for recruitment automation. While this research flipped recruitment automation, by ranking jobs for a given resume, the results from this research are still relevant to recruitment automation. This research highlights how the TF-IDF model often struggled to generalise terms beyond surface-level similarity, leading to weak results in a variety of domains. In contrast, SBERT was capable of understanding words can be on job descriptions as another word, with many domains using synonyms for required skills or tools. The key distinction between these two models, with the results of the evaluations from this study in mind, leads this study to recommend the use of SBERT for resume-job matching tasks.

This study highlighted that while SBERT was capable of more consistent, and accurate results, this came with a longer runtime. With a runtime multiple times longer than the TF-IDF model, SBERT might not be the one-size-fits-all solution for large datasets. While TF-IDF was much faster, the memory usage was much higher, unlike SBERT's lower and consistent memory usage. The computational cost from both models is the key element of highlighting trade-offs between the two models.

In conclusion, this research supports that for the task of resume-job matching, SBERT is the more effective model. This answers the research question by concluding that semantic-based embedding models are not only capable of outperforming traditional methods on accuracy metrics, but also in providing jobs that are aligned with human interpretation. This research highlights SBERT as the stronger model for this task in large-scale operations where data is diverse and results need to be accurate, consistent and align with human interpretation.

6.2 Limitations of Study

While the findings of this study are strong and cover a variety of areas, several limitations must be noted.

A. Lack of Fine-Tuning

The SBERT model was used as-is, without any fine-tuning. New alternative models such as conSultantBERT and CareerBERT, would have had different results with their fine-tuning being domain specific, but they were outside the scope of this study.

B. Qualitative Sample Size

The qualitative evaluation within this study consisted of five participants. This evaluation provided valuable insight, highlighting the differences in results from each model, from a human perspective. Additionally, the participants had a background in hiring and had the required information to effectively rank the outputs. The key limitation in this section is the small sample size. A large sample size would strengthen the findings and likely further add to the study.

C. Dataset Size & Diversity

The job postings dataset used for this research does not fully capture every role and domain in hiring to the same extent of large recruitment platforms like LinkedIn. While the chosen dataset consists of information from LinkedIn from 2023-2024, this is not the same as having a dataset the same as LinkedIns. Real resume-job matching involves millions of postings across diverse industries, leading to this being a limitation within this study. A larger dataset is likely to provide different results, highlighting each model's strengths and limitations in certain domains.

6.3 Future Works

Future work of resume-job matching should take the following suggestions into consideration.

A. Fine-Tuning Models on Domain Specific Data

Using fine-tuned models on domain specific job datasets could allow the model to learn hiring-specific semantics. This is likely to increase the model's performance in domains with diverse jobs where skills, tools and experience can fluctuate per job listing. While SBERT was capable of effective performance in domains like this (*Graphic Designer)*, a fine-tuned model per domain would outperform the model used in this study.

B. Larger Qualitative Evaluation

A larger qualitative evaluation with participants having experience in human resources, recruitment or domain experts, could provide stronger qualitative results. While the small sample size of five participants was sufficient for this study, larger studies surrounding which models to implement into organisations could utilize a larger test group for more accurate results.

6.4 Final Remarks

This research strongly suggests that transformer-based semantic similarity models are capable of results, a traditional TF-IDF model can't replicate. The chosen transformer-based model, SBERT, was capable of displaying meaningful improvements over traditional keyword matching for resume-job matching. While evaluation results from TF-IDF were weaker than SBERT, it still remains useful in fast-filtering contexts and has shown to be effective in domains with sufficient lexical overlap. In domains where nuance, context and deeper meanings must be extracted, SBERT is the ideal model. The outcomes of this study add to the research of transformer-based models, highlighting the strengths of the SBERT model at resume-job matching. The following is a link to a presentation further discussing the research.

References

1. Bevara, R. V. K., Mannuru, N. R., Karedla, S. P., Lund, B., Xiao, T., Pasem, H., Dronavalli, S. C. and Rupeshkumar, S. (2025), 'Resume2vec: Transforming applicant tracking systems with intelligent resume embeddings for precise candidate matching', Electronics 14(4), 794. URL: https://www.mdpi.com/2079-9292/14/4/794
2. Bhatia, V., Rawat, P., Shah, R. R. and Kumar, A. (2019), 'End-to-end resume parsing and finding candidates for a job description using bert', arXiv preprint arXiv:1910.03089 . URL: https://arxiv.org/pdf/1910.03089
3. Bhawal, S. (2025), 'Resume dataset'. URL: https://www.kaggle.com/datasets/snehaanbhawal/resume-dataset
4. Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019), BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. URL: https://arxiv.org/pdf/1810.04805
5. Elsafty, A., Riedl, M. and Biemann, C. (2018), Document-based Recommender System for Job Postings using Dense Representations. URL: https://aclanthology.org/N18-3027.pdf
6. Gillioz, A., Casas, J., Mugellini, E. and Khaled, O. A. (2020), 'Overview of the transformer-based models for nlp tasks', Proceedings of the 2020 Federated Conference on Computer Science and Information Systems 21.
7. Jirjees, A. K., Ahmed, A. M., Abdulla, A. A., Lu, J., Noori, E. M., Kareem, R. N., Hassan, B. A., Veisi, H. and Rashid, T. A. (2025), 'Machine learning for recruitment: Analyzing job matching algorithms', TechRxiv . Preprint. Not peer reviewed. Posted under CC-BY 4.0 License. URL: https://doi.org/10.36227/techrxiv.174240031.12397312.v1
8. Khan, S., Shahi, A., Alam, A., Bafna, A. and Therese, S. (2025), 'Comparison of models for resume-jd matching: Bert, gemini, and llama 3.1', Journal of Computer Engineering 27(2), 1–10. URL: https://www.iosrjournals.org/iosr-jce/papers/Vol27-issue2/Ser-5/A2702050110.pdf
9. Koneru, A. and Yu Zou, Z. (2025), 'Linkedin job postings (2023 - 2024)'. URL: https://www.kaggle.com/datasets/arshkon/linkedin-job-postings
10. Lavi, D., Medentsiy, V. and Graus, D. (2021), conSultantBERT: Fine-tuned Siamese Sentence-BERT for Matching Jobs and Job Seekers. URL: https://arxiv.org/pdf/2109.06501
11. Nigam, A., Roy, A., Saxena, A. and Singh, H. (2019), 'Job recommendation: Leveraging progression of job applications', arXiv preprint arXiv:1905.13136 . URL: https://arxiv.org/pdf/1905.13136
12. Peffers, K., Tuunanen, T., Rothenberger, M.A. and Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, 24(3), pp.45–77. doi:https://doi.org/10.2753/mis0742-1222240302.

13. Qaiser, S. and Ali, R. (2018), 'Text mining: Use of tf-idf to examine the relevance of words to documents', International Journal of Computer Applications 181(1), 25–29.

14. Reimers, N. and Gurevych, I. (2019), Sentence-bert: Sentence embeddings using siamese bert-networks, in 'Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing', As- sociation for Computational Linguistics, pp. 3982–3992. URL: https://arxiv.org/abs/1908.10084

15. Rosenberger, J., Wolfrum, L., Weinzierl, S., Kraus, M. and Zschech, P. (2025), 'Careerbert: Matching resumes to esco jobs in a shared embedding space for generic job recommendations', Expert Systems with Applications 275, 127043. URL: https://www.sciencedirect.com/science/article/pii/S0957417425006657

16. Yu, X., Zhang, J. and Yu, Z. (2024), 'Confit: Improving resume-job matching using data augmentation and contrastive learning'. URL: https://arxiv.org/abs/2401.16349

17. Linkedin Talent Solutions (2024). *The Future of Recruiting 2024 2 2*. [online] Available at: https://business.linkedin.com/content/dam/me/business/en-us/talent-solutions/resources/pdfs/future-of-recruiting-2024.pdf.

18. Sasi, K.P. (2024). Impact of AI in Recruitment and Talent Acquisition. *Human resource and leadership journal*, 9(3), pp.78–83. doi: https://doi.org/10.47941/hrlj.2117.

19. Sartor, G. and Lagioia, F. (2020). *The impact of the general data protection regulation on artificial intelligence*. *European Parliament eBooks*, European Parliament, pp.1–84. doi:https://doi.org/10.2861/293.

20. Mujtaba, D.F. and Mahapatra , N.R. (2019). *Fairness in AI-Driven Recruitment: Challenges, Metrics, Methods, and Future Directions*. [online] Arxiv.org. Available at: https://arxiv.org/html/2405.19699v3

21. Rigotti, C. and Fosch-Villaronga, E. (2024). Fairness, AI & Recruitment. *Computer law and security report/Computer law & security report*, [online] 53(1), pp.105966–105966. doi:https://doi.org/10.1016/j.clsr.2024.105966.

22. SBERT (2025). *Retrieve & Re-Rank — Sentence Transformers documentation*. [online] Sbert.net. Available at: https://sbert.net/examples/sentence_transformer/applications/retrieve_rerank/README.html? [Accessed 17 Oct. 2025].

23. Walsh, H.S. and Andrade, S.R. (2022). *Semantic Search with Sentence-BERT for Design Information Retrieval*. [online] Nasa.gov. Available at: https://ntrs.nasa.gov/citations/20220002323 [Accessed 17 Oct. 2025].