



Title: Intelligent Job Recommendation System for Job Seekers Using Machine learning.

Prepared By :

Ahmad Munir Mzayek

ID No.: [243001713](#)

Submitted in fulfillment of the requirements for the degree
Master of Science in Artificial Intelligence

At

College of Informatics

Midocean University

UAE

Supervisor Name: Dr. Ashraf Saeed

College of Informatics, Midocean University, Comoros

[[2025-3rd-semester](#)]

Acknowledgements

In the name of God, Most Gracious, Most Merciful.

All praise and thanks are due to Allah for His grace and for granting me the strength and will to complete this project successfully.

I would also like to express my sincere appreciation to Dr. Ashraf for the priceless guidance offered during this journey. I also wish to express my profound gratitude to my family for their unlimited support from start to finish, their kindness I will never forget in my life.

God willing, I will work hard to devote what I have learned to doing good and training and educating people on artificial intelligence because it is the future in our practical and academic life, also I would like to dedicate this accomplishment to my beloved Syria, which is always in my heart, and the Saudi Arabian Kingdom, my second home, for the opportunities and generosity I received here.

Finally,

"اللهم علمنا ما ينفعنا، وانفعنا بما علمتنا، وزدنا علما "سُبْحَانَكَ لَا عِلْمَ لَنَا إِلَّا مَا عَلَّمْتَنَا إِنَّكَ أَنْتَ الْعَلِيمُ الْحَكِيمُ"

Abstract

The recent rise in online portals for recruitment has caused the process of linking the job seekers to the appropriate opportunities to become complex. The current method of recommending jobs to applicants has been mainly based on keyword matching approaches, which ignore the semantics attaching the CVs to the work description, especially where multiple languages are involved.

This thesis aims at the design of an intelligent job recommendation system using semantic text embeddings and machine learning algorithms. This system tests the job-resume matching on various experimental tasks, such as on CSV-based structured datasets as well as PDF resumes in English and Arabic languages. Four semantic text embeddings are being compared: Word2Vec, TF-IDF Weighted Word2Vec, Document2Vec, and Sentence-BERT. Cosine similarity measure is being used for the ranking metric, along with the Top-1 selection strategy.

In the context of Arabic resumes, the performance of several machine translation paradigms, including Google Translate, mTranslate, NLLB, and M2M100, is also explored to determine the effect of the quality of translation being used on the results of the semantic matching. The experiment confirms the superiority of the proposed SBERT approach over the existing statistical and document embedding paradigms for all the data sets. The results of the experiment also indicate the effect of translation quality of the Arabic resumes.

The results validate that deep contextual embeddings offer far greater robustness and semantic matching capabilities to job recommendation systems. In addition, it is observed that the size of the dataset, recommendation approach, and multilingual representation play critical roles in boosting the performance. This work makes a significant addition to a scalable and extendable framework for efficient intelligent job recommendation systems, addressing real-world issues involving multilingual recruitment.

The experiment results show that Sentence-BERT achieves the most reliable and robust semantic matching capabilities when performing recruitment tasks on structural, unstructured, and multilingual recruitment.

Contents

Acknowledgements	2
Abstract	3
Chapter 1: Introduction	7
1.1 Problem Definition	7
1.2 Thesis Objectives	7
1.3 Thesis Contribution	8
1.4 Thesis Organization	8
Chapter 2: Overview	9
2.1 Introduction	9
2.2 Job Recommendation Systems	9
2.3 Natural Language Processing in Recruitment Systems	9
2.4 Text Representation and Embedding Models	9
2.5 Multilingual Processing and Translation	10
2.6 Similarity Measurement	10
2.7 Chapter Summary	10
Chapter 3: Related Work	11
3.1 Introduction	11
3.2 Content-Based Job Recommendation Systems	11
3.3 Semantic Embedding-Based Approaches	11
3.4 Resume Matching and Ranking Systems	11
3.5 Multilingual and Translation-Based Systems	12
3.6 Evaluation of Recommendation Systems	12
3.7 Public Datasets for Recruitment Research	12
3.8 Limitations of Existing Approaches	12
3.9 Chapter Summary	12
Chapter 4: Proposed Approach	13
4.1 Introduction	13
4.2 System Architecture	13
4.3 Experimental Scenarios	13

4.3.1 Excel-Based English Resume Matching (CSV)	13
4.3.2 PDF-Based English Resume Matching	13
4.3.3 PDF-Based Arabic Resume Matching with Translation	13
4.4 Text Preprocessing	13
4.5 Text Embedding Models	14
4.5.1 Simple Word2Vec	14
4.5.2 TF-IDF Weighted Word2Vec	14
4.5.3 Doc2Vec	15
4.5.4 Sentence-BERT (SBERT)	15
4.5.5 Cosine similarity	15
4.6 Similarity Computation and Top-1 Selection	16
4.7 Chapter Summary	16
4.8 How the recommendation system works	16
Chapter 5: Experimental Results and Discussion	17
5.1 English Dataset – CSV (Excel-Based Experiment)	17
5.2 English PDF Resumes Results	18
5.2 Arabic PDF Resumes Results (Google + mTranslate)	19
5.3 Arabic PDF Resumes Results (NLLB + M2M100)	20
5.4 Interpretation – General Observations and Improvement Directions	21
5.4 Discussion	21
Chapter 6: Conclusion and Future Work	22
6.1 Conclusion	22
6.2 Future Work	22
Chapter 7: References	23

Table 1- English Dataset	17
Table 2- English Results	18
Table 3- English PDF Resumes Results	18
Table 4- Arabic PDF Resumes Results (Google)	19
Table 5- Arabic PDF Resumes Results (mTranslate)	19
Table 6- Arabic PDF Resumes Results (NLLB)	20
Table 7- Arabic PDF Resumes Results (M2M100)	20
Figure 1- processing procedure pipeline	16
Equation 1- Word2vec	14
Equation 2- Cosine similarity	14
Equation 3- TF-IDF	14
Equation 4- Doc2vec	15
Equation 5- SBERT	15
Equation 6- Cosine similarity	15

Chapter 1: Introduction

1.1 Problem Definition

This rapid growth in online job portals has resulted in an overwhelming number of job listings as well as resumes from the desired candidates. Despite the increased accessibility to job opportunities, there lie many difficulties in properly mapping the desired candidates with the appropriate job listings. This is because most current job recommendation systems use a keyword-based approach for mapping the resumes with the required jobs without understanding the semantic connections between the two aspects.

This challenge becomes more complicated in a multilingual environment where the job descriptions and CVs happen to be in different languages, say English and Arabic. The imperfection in results might also occur even when the use of translations has been adopted and implemented due to challenges associated with the different languages, translations, and situations. Therefore, appropriate job vacancies may fall through the sieve, while inappropriate suggestions may also be provided to the job seekers.

Also, most systems at present do not consider conceptual similarities between resumes and vacancy descriptions at the document level, resulting in mismatches between matched resumes and vacancy descriptions that share similar meanings but use different words. Besides, most of the current systems do not consider conceptual similarities between resumes and vacancy descriptions at the document level; this resulted in mismatches between resumes and descriptions that shared the same meanings but with different words used. The shortcomings mentioned above emphasize the importance of an intelligent resume recommendation system that is semantically intelligent, multilingual, and a good similarity matcher.

1.2 Thesis Objectives

The main aim of the thesis is to devise and test the effectiveness of an intelligent job recommendation system that enhances the relevance of matching job seekers with potential job opportunities through machine learning and natural language processing approaches. The specific objectives include the following:

- To build a job recommendation system for resumes written in English and Arabic.
- For the evaluation of the proposed method of computing the semantic relationships between the input and the two contexts within the framework.
- To compare traditional statistical models and deep learning-based semantic models.
- For probing the effect of translation techniques on resume matching done in Arabic language.

- The embedding representation most effectively applicable to the semantic task of recommending jobs.

1.3 Thesis Contribution

This thesis proposal discusses the design of an intelligent job recommendation system end-to-end framework based on the processing of job descriptions and resumes through the use of data preprocessing, embedding, and similarity calculation processes. An experimental investigation based on the use of four different embedding models is carried out: Word2Vec, TF-IDF-weighted Word2Vec, Doc2Vec, and Sentence BERT (SBERT).

This study further investigates the issues associated with multilingual job recommendations by considering the performance comparison of varied translation models on Arabic résumés. Experiments validate the results on the superior performance of deep context embeds than statistical models.

1.4 Thesis Organization

This thesis has six chapters. In chapter 1, introduction to the research problem, goals, and contributions are mentioned. The theoretical basis of this research has been described in chapter 2. Related research has been reviewed in chapter 3. My proposal on research methodology has been described in chapter 4. The experimental analysis and outcome of this research has been mentioned in chapter 5. Conclusion of this research thesis has been described in chapter 6.

Chapter 2: Overview

2.1 Introduction

Artificial Intelligence and Natural Language Processing have contributed immensely to new-era recruitment technology by facilitating the processing of unstructured textual data in resumes or employment description posts. Job recommendation systems also aim to alleviate the burden of candidates by automatically suggesting them employment opportunities based on their requirements.

This chapter introduces or explains different concepts that are relevant to the research, namely job recommendation systems, text embedding, multilingual text, and similarity measurement.

2.2 Job Recommendation Systems

Jobs recommender systems are specific recommender systems that are aimed at linking the profiles of potential employees with the available job postings. Traditional job recommender systems have utilized rule-based systems or keyword-based matching algorithms, which are computationally effective but semantically poor.

Current approaches involve machine learning as well as content-based filtering, in which resumes and job descriptions can be expressed as numerical vectors. This makes it possible to rank the job postings by resume profile on the basis of similarities.

2.3 Natural Language Processing in Recruitment Systems

The application of NLP processes is crucial when handling the unstructured hiring data. The most common preprocessing tasks involved are text retrieval, normalization, tokenization, stop words elimination, and noise reduction. These processes help to boost the quality of data.

Semantic modeling is highly useful when it comes to job recruitment, as it allows a measure of conceptual similarity to be determined.

2.4 Text Representation and Embedding Models

The drawback with traditional models such as Bag of Words and TF-IDF is sparsity and a shallow representation of meaning. Embeddings such as Word2Vec and Doc2Vec solve this by learning a dense representation.

More recently, models such as Sentence-BERT are designed as transformers and are capable of learning contextualized representations that are specifically optimized for semantic similarity-based tasks.

2.5 Multilingual Processing and Translation

Multi-lingual job recommendation platforms rely on machine translation to create a unified space between resumes and job postings in a single language. However, this would have a major effect on semantic consistency for a morphology-rich language such as Arabic.

2.6 Similarity Measurement

The cosine similarity function has been widely used in computing semantic similarities from vector representations because it has normalizing properties.

2.7 Chapter Summary

This chapter has described the underlying theories which allow intelligent job recommendation systems and gave a rationale for building a methodological tool.

Chapter 3: Related Work

3.1 Introduction

Job recommendation systems have received considerable interest in the literature because of the increasing popularity of recruitment websites. The primary goal of designing these systems is to ease the search process for candidates while increasing the accuracy of suggestions. This chapter presents an overview of the literature pertaining to job recommendation systems, semantic representation of text, as well as multilingual recruitment tools.

3.2 Content-Based Job Recommendation Systems

Initial job recommendation systems were primarily based on the use of content-based filtering, revolving around the idea of comparing CVs and job descriptions based on their similarity in text. The idea of a job recommendation system based on both content-based and collaborative filtering was proposed by Dhameliya and Desai in [1], stating that keyword-based approaches give good results but lack semantic intelligence.

To overcome this limitation, Mulay et al. [2] proposed hybrid filtering techniques that integrated content-based approaches with collaborative filtering. These hybrid techniques were found to be more accurate in terms of recommendation, though they were highly dependent on historical user interaction data.

3.3 Semantic Embedding-Based Approaches

Representation learning has progressed, and as a result, embedding models are applied to job recommendation systems. Reimers and Gurevych's work "Sentence-BERT: Sentence Embeddings using BERT" introduced a Sentence-BERT (SBERT) embedding that produced a fixed-size embedding for a given set of sentences to perform semantic similarity. The system outperformed existing models that used embeddings.

Thite [4] used SBERT model on resume ranking tasks and proved that SBERT could effectively learn contextual similarity between resumes and job descriptions. Vanetik [6] also declared that, on job vacancy ranking tasks, sentence embeddings work even better than statistical models.

3.4 Resume Matching and Ranking Systems

Recent works include intelligent matching systems on resumes. Sagar et al. proposed an intelligent resume matching system using machine learning techniques in [5], pointing out that traditional keyword-based matching systems lack efficiency and that there is a need to assess semantic similarity.

Additionally, the work of Singla and Verma [9] showed the importance of embedding models in improving the accuracy of job recommendation, especially when used in conjunction with machine learning classifiers.

3.5 Multilingual and Translation-Based Systems

The multilingual job recommendation systems pose an extra level of complexity due to the variance in language. Ertuğrul and Bitirim in [10] performed a systematic literature study, which found translation quality to be a challenge in multilingual job recommendation systems. Even using Neural Machine Translation, semantic distortion is a known prevailing problem.

3.6 Evaluation of Recommendation Systems

Cremonesi et al. [11] analyzed the assessment of Top-N recommendation tasks and the significance of the precision of ranking accuracy metrics. Their results provide an indication of the effectiveness of the Top-1 and Top-N ranking approaches for the evaluation of recommender systems.

3.7 Public Datasets for Recruitment Research

Some public datasets supporting research in recruitment were proposed. Elabady [12] published the Wuzzuf job listings dataset, whilst Dutta [13], Koneru and Zou [14], and Bhawal [15] published the resume and job listings dataset for a job recommendation system.

3.8 Limitations of Existing Approaches

Notably, although a great deal has been accomplished by preceding research on job recommendation systems, a majority of existing models are solely based on keyword matching or utilize limited semantic structures. Differently, instead of solely emphasizing a specific dimension as done by existing research, the proposed work within this thesis contributes to a comprehensive comparison between statistical models, document models, and contextual embedding models on both structured and unstructured data, as well as on a multilingual perspective.

3.9 Chapter Summary

The current chapter has reviewed the existing work in the context of the recommendation systems for jobs, methods involving semantic embeddings, processing multiple languages, and existing methods for the evaluation process. The drawbacks in the existing methods form the basis for the proposed methodology in the thesis.

Chapter 4: Proposed Approach

4.1 Introduction

In this chapter, the proposed approach for building an intelligent job recommendation system that can cope with multilingual resumes and perform semantic job matching will be introduced. The system has been designed as a structured pipeline that can support different experimental scenarios, either using Excel-based datasets or actual PDF resumes.

4.2 System Architecture

The proposed design is based on a modular approach with steps involved in the design including input, preprocessing, feature extraction, similarity calculation, and recommendation ranking.

4.3 Experimental Scenarios

4.3.1 Excel-Based English Resume Matching (CSV)

With the initial experiment design, structured data inputs were employed to provide baseline testing. For this experiment, the resume data was the Resume 2.csv file, while the training data consisted of the training_data2.csv file.

4.3.2 PDF-Based English Resume Matching

The second scenario used actual resumes in PDF format that are common in England. The text extraction was carried out using PDF processing tools, followed by embedding and preprocessing. This scenario aimed to test the strength of the model in realistic conditions of unstructured text.

4.3.3 PDF-Based Arabic Resume Matching with Translation

The third scenario focused on bilingual issues by evaluating the translations of resumes written in Arabic but processed as if written in English by employing four different translation models: Google Translate, mTranslate, NLLB, and M2M100.

4.4 Text Preprocessing

The preprocessing operations involved performing tasks such as text normalization, tokenization, stop word removal, as well as noise removal. For the Arabic resumes in particular, the operation of translation was carried out before the preprocessing phase in order to have a uniform semantic space.

4.5 Text Embedding Models

Four embedding models were evaluated

4.5.1 Simple Word2Vec

Word2Vec represents every word with a dense numerical vector in a continuous vector space. A resume or job description was represented by averaging the vectors of all words contained within the document:

$$\vec{V}_{doc} = \frac{1}{n} \sum_{i=1}^n \vec{v}_i$$

Equation 1- Word2vec

where $\vec{V}_{doc}$ is the document vector (resume or job description), n is the number of words in the document (after preprocessing), and $\vec{v}_i$ is the Word2Vec embedding vector of the i -th word.

Cosine similarity between resume and job vectors was computed as:

$$Similarity(\vec{V}_r, \vec{V}_j) = \frac{\vec{V}_r \cdot \vec{V}_j}{\|\vec{V}_r\| \|\vec{V}_j\|}$$

Equation 2- Cosine similarity

where $\vec{V}_r$ represents the resume vector and $\vec{V}_j$ represents the job description vector.

While this way is computationally efficient, it doesn't account for word importance or deep contextual meaning.

4.5.2 TF-IDF Weighted Word2Vec

In this model, word vectors were weighted with their TF-IDF score to accentuate the important terms:

$$\vec{V}_{doc} = \frac{\sum_{i=1}^n TFIDF(w_i) \cdot \vec{v}_i}{\sum_{i=1}^n TFIDF(w_i)}$$

Equation 3- TF-IDF

Where w_i is the i -th word, $TFIDF(w_i)$ is the TF-IDF weight (importance) of that word, and $\vec{v}_i$ is its Word2Vec embedding; the denominator normalizes the weighted sum.

This approach improves discriminative power but still lacks contextual understanding.

4.5.3 Doc2Vec

The Doc2Vec algorithm obtains a unique vector representation of each document in

$$\vec{D} = f(\vec{w}_1, \vec{w}_2, \dots, \vec{w}_n)$$

Equation 4- Doc2vec

Where $\vec{D}$ is the learned document embedding, $f(\cdot)$ represents the Doc2Vec model (a learned function), and $\vec{w}_1, \dots, \vec{w}_n$ denote the word representations/inputs from the document used by the model to infer the document vector.

It is highly sensitive to the size of the dataset and will generate unstable embeddings on smaller datasets.

4.5.4 Sentence-BERT (SBERT)

It generates embeddings fine-tuned for semantic similarity. The all-mpnet-base-v2 model was used:

$$\vec{E} = \text{Transformer}(\text{text})$$

Equation 5- SBERT

Where $\vec{E}$ is the contextual embedding vector produced by the SBERT model (e.g., all-mpnet-base-v2), and text is the full resume or job description input.

SBERT catch deep contextual semantics and crop sure similarity distributions.

4.5.5 Cosine similarity

Cosine similarity is calculated between each resume and the job description.

This measure computes the cosine of the angle between two vector representations and is defined as:

$$\text{Similarity}(V_r, V_j) = \frac{V_r \cdot V_j}{\|V_r\| \|V_j\|}$$

Equation 6- Cosine similarity

where $\vec{V}_r$ represents the resume vector and $\vec{V}_j$ represents the job description vector. Similarity scores were then used to rank the jobs for each resume. The cosine similarity values also show which job postings are more relevant; therefore, the job postings with higher values ranked on top.

4.6 Similarity Computation and Top-1 Selection

Cosine similarity was used to calculate the semantic similarity between the job posting and the resume vectors. The job with the highest similarity score for each resume was selected as the Top-1 job. The job selection was done without any human evaluation.

Background on the Top-1 evaluation was required. This approach was adopted in the study to ensure that the evaluation was undertaken in an objective and comparable manner. Moreover, the focus on the best-ranked recommendation overlooked the impact that the selection of various recommendations may have on the evaluation. This approach ensures that the evaluation is strictly undertaken. Despite the potential offered by Top-N recommendations, the approach adopted was the best strategy for evaluation.

4.7 Chapter Summary

This chapter explained the methodological approach, experimental conditions, and technological aspects of the proposed job recommendation system. The proposed framework is designed to make a fair comparison among machine learning models and datasets.

4.8 How the recommendation system works

A simple 7 steps pipeline was design to process any resume and find the best job match

“Input – Text processing and cleaning – 4 embedding model – training dataset – similarity computation – best match – final job recommendation output “

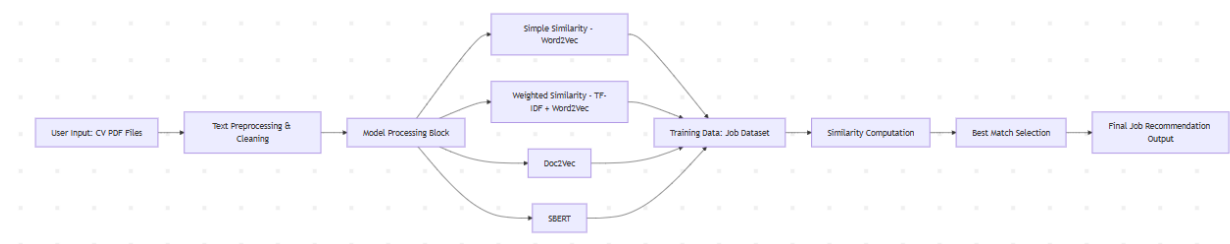


Figure 1- processing procedure pipeline

Chapter 5: Experimental Results and Discussion

5.1 English Dataset – CSV (Excel-Based Experiment)

The following experimental scenario will consider structured datasets to assess the performance of semantic job recommendation models in controlled conditions. Resume 2.csv contains candidate resume profiles, and training_data2.csv comprises job descriptions. Recommendation models can be provided with these kinds of datasets as input, and the results will be noted in the Top1_Results_4Models_With_Original_Resume_ID file.

For each resume, the recommendation procedure entailed calculating cosine similarity scores between the resume representation and all job descriptions. The job with the highest similarity score was selected as the Top-1 recommendation. This selection process was performed automatically-without any human intervention-to ensure objective and reproducible evaluation.

The evaluation was based on the semantic alignment between the predicted Top-1 job and the original profile of the resume. The assessment was based on three criteria: relevance of the job domain, similarity in job title, and overall semantic consistency between candidate skills and the requirements of the job.

The comparative analysis actually confirms that the deep contextual embedding models far outperform the traditional statistical ones in structured job recommendation tasks. SBERT had the highest consistency, the lowest semantic mismatch rate, and the highest alignment intensity of resumes versus job postings. Such findings justify SBERT as the most suitable model for job recommendation systems operating on structured datasets.

Dataset	Type	Number of Records	Description
Resume 2.csv	Resumes	~2800	Candidate resume profiles
training_data2.csv	Job	~1900	Job descriptions
Top1 Results	Output	100	Top-1 predictions per model

Table 1- English Dataset

Model	Domain Match	Title Match	Overall Consistency
Word2Vec	Disagree	Disagree	Weak
TF-IDF Word2Vec	Weakly Agree	Disagree	Moderate
Doc2Vec	Weakly Agree	Weakly Agree	Unstable
SBERT	Strongly Agree	Strongly Agree	Strongly Agree

Table 2- English Results

5.2 English PDF Resumes Results

In total of 30 real resumes:

	Simple	Weighted	Doc2Vec	SBERT
Strongly Agree	47%	43%	43%	63%
Weakly Agree	20%	13%	20%	23%
Disagree	33%	43%	37%	13%

Table 3- English PDF Resumes Results

Interpretation

The test results from the English language PDF resumes indicate a clear superiority of the semantic embedding models over the conventional statistical models. Of all the models tested, the one that performed best in identifying highly relevant results was the Sentence-BERT (SBERT) model, which recorded the highest percentage of Strongly Agree (63%) with the lowest percentage of Disagree (13%). This result indicates a high ability to recognize the resume-descriptor semantic relations by the SBERT algorithm.

The robust performance of word2vec models in the English environment can be considered to be due to the absence of translation noise and a stable vocabulary. The word2vec models still fare poorly in comparison to SBERT because of their limitations in capturing deeper meaning in a context. The performance of models in the English PDF case suggests that SBERT is the best approach to adopt for monolingual recommendation systems for the job.

5.2 Arabic PDF Resumes Results (Google + mTranslate)

In total of 33 real + generated resumes:

<u>Google</u>	Simple ID	Weighted ID	Doc2Vec ID	SBERT ID
Strongly Agree	12%	0%	21%	30%
Weakly Agree	3%	6%	15%	3%
Disagree	85%	94%	64%	67%

Table 4- Arabic PDF Resumes Results (Google)

<u>mtranslate</u>	Simple ID	Weighted ID	Doc2Vec ID	SBERT ID
Strongly Agree	12%	0%	24%	30%
Weakly Agree	3%	6%	12%	3%
Disagree	85%	94%	64%	67%

Table 5- Arabic PDF Resumes Results (mTranslate)

Interpretation

The obtained results for the translated Arabic resumes using Google Translate and mTranslate indicate a remarkably lower performance regarding the accuracy of matches in comparison to the results of the English dataset. However, the performance of the SBERT model is always the best among other models, with the highest rate of Strongly Agree of 30%.

By comparison, the Simple Word2Vec model as well as the TF-IDF weighted Word2Vec model had very high rates of mismatch, crossing the 90% Disagree mark in some instances. This was because these models depended entirely on word overlap as their primary means of retaining meaning despite translation. Although the Doc2Vec model showed some improvement in this aspect, it was not very stable.

These results verify that machine translation does indeed introduce semantic noise that affects the quality of job matching, and that deep contextual embeddings perform better in overcoming the mentioned problem, though not eliminating it.

5.3 Arabic PDF Resumes Results (NLLB + M2M100)

In total of 33 real + generated resumes:

<u>NLLB</u>	Simple ID	Weighted ID	Doc2Vec ID	SBERT ID
Strongly Agree	3%	3%	18%	27%
Weakly Agree	0%	3%	6%	6%
Disagree	97%	94%	76%	67%

Table 6- Arabic PDF Resumes Results (NLLB)

<u>M2M100</u>	Simple ID	Weighted ID	Doc2Vec ID	SBERT ID
Strongly Agree	15%	3%	9%	15%
Weakly Agree	0%	18%	0%	0%
Disagree	85%	79%	91%	85%

Table 7- Arabic PDF Resumes Results (M2M100)

Interpretation

The above research results using the advanced neural translation models for semantic matching further reinforce the “hourglass”-shaped challenge curve for the task of semantic matching for the many languages involved in the project here. While the results using the translation models show a slight improvement over the competition for Google and mTranslate when using SBERT, the task of semantic transfer remained a challenge for the different models used. SBERT again performed better when compared with the other embeddings used in the task for the highest rate of Strongly Agree responses given by the respondents.

The fluctuating performance pattern identified in the case of M2M100 and the high rates of mismatch suggest that the improvement in the quality of the translations is not sufficient to ensure the semantic alignment of the recommendations. These observations illustrate the fact that multilingual translation solutions based upon translation are fundamentally constrained and the semantic deterioration continues to persevere despite the presence of the most advanced translation networks.

5.4 Interpretation – General Observations and Improvement Directions

As far as overall experimental findings are concerned, it has been proven that SBERT is the most robust and reliable embedding model on both English and Arabic databases. However, there seems to be some limitation in Arabic environments, which indicates that following one Top-1 suggestion might not be sufficient in a multilingual environment. Instead, following suggestions on Top-3 or even Top-N suggestions would help enhance the chances of getting relevant job matches in environments with translation uncertainty. In addition, a larger and more diverse pool of resumes is expected to enhance model generalization and position stability for ranking models like Doc2Vec in particular. Future work on this project may include enhancing job recommendations for multi-languages by employing native models of semantic embeddings of Arabic and/or other languages to lower reliance on machine translation models and enhance job recommendation results.

5.4 Discussion

There is a discernible performance discrepancy when comparing the organized experiments involving CSV files and those involving PDF files. The optimal performance obtained in the CSV environment can be attributed to the fact that it is a noise-free environment, free from semantic noise that might have resulted from processing resumes involving machine translations. In contrast, PDF resumes, especially when dealing with a multilingual environment, add noise to semantic performance. Again, it is clear that although data format affects performance, it is a secondary determinant compared to embedding strategy.

Putting the structured CSV-based experiments together with the results obtained from both English and Arabic PDF resumes indicates a similar pattern regarding performance. In all experimental scenarios, Sentence-BERT performed better than the other embedding models, hence confirming its robustness across different data formats and linguistic settings.

The Excel-based experiments were done in a tightly controlled environment where noise introduced by PDF text extraction and machine translation were eliminated. In this context, indeed, the best semantic matching was achieved with SBERT, further convincing that contextual embeddings are needed for an accurate job recommendation system. Low comparative performance in the Arabic PDF experiments underlines the challenges introduced by semantic loss due to translations. These results together show that model architecture makes much more difference than the data format. In addition, the dataset size and extension of recommendations from Top-1 to Top-3 are alone expected to result in increased performance; the embedding model chosen remains the most preponderant factor influencing recommendation accuracy.

Chapter 6: Conclusion and Future Work

6.1 Conclusion

In this thesis work, an intelligent job recommendation system based on semantic text embedding models and machine learning algorithms has been proposed and tested. A significant amount of experimentation work has been conducted on both structured CSV datasets as well as real-world PDF resumes to evaluate the performance of the models in a controlled as well as real-world setting with encouraging results showing the superiority of deep contextual embedding models over traditional models.

For all tested cases, the performance of Sentence-BERT was outstanding on both the English and Arabic datasets in identifying the relationships between resumes and job descriptions. The results also indicate that the multilingual translation task of recommending a job remains challenging, even with highly accurate translation models, owing to the effects of semantic loss incurred during machine translation. The importance of a language-aware approach to semantic representation in enhancing the accuracy of recommendation in a multilingual setting is therefore a critical lesson learned from this experimentation.

6.2 Future Work

In the future, the strategy could be extended from Top-1 recommendation to Top-N recommendation to achieve more robust matching, especially in the context where there is ambiguity in translation. On the other hand, expansion in the diversity of the dataset should improve the ranking model, especially the document-level embedding model like Doc2Vec.

Other areas that could be further advanced are incorporating methods of collaborative filtering to better personalization, and lastly, utilizing native models of semantic embedding in Arabic to decrease dependence on MT. Finally, incorporating methods of Explainable AI may improve understandability and enable a better level of trust in intelligent job recommendation tools.

Chapter 7: References

- 1- Juhi Dhameliya, Nikita Desai, Job Recommendation System using Content and Collaborative Filtering based Techniques, International Journal of Scientific Research in Engineering and Management (IJSREM), Volume 9 - 2019, Issue-3, September 2019.
<https://www.ijsce.org/wp-content/uploads/papers/v9i3/C3266099319.pdf>
- 2- Aneesh Mulay, Shriyash Sutar, Jiten Patel, Aditi Chhabria, Snehal Mumbaikar, Job Recommendation System Using Hybrid Filtering, researchgate, Volume 44 – 2022, Issue 02002, January 2022.
https://www.researchgate.net/publication/360418721_Job_Recommendation_System_Using_Hybrid_Filtering
- 3- Nils Reimers, Iryna Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP 2019), 2019.
<https://arxiv.org/abs/1908.10084>
- 4- Abhishek Thite, Resume Ranking for a Job Description Deriving Similarity of Representational Embedding Using Sentence-BERT, researchgate, Liverpool John Moores University (Master Thesis), 2022.
https://www.researchgate.net/publication/383953314_RESUME_RANKING_FOR_A_JOB_DESCRIPTION_DERIVING_SIMILARITY_OF_REPRESENTATIONAL_EMBEDDING_USING_SENTENCE_BERT
- 5- Dr. Y. Sagar, K. S. G. Prakeerth, G. Abhiram, N. Mahideep, M. Poojith, Intelligent Resume Matching System, International Journal for Research in Applied Science & Engineering Technology (IJRASET), Volume 13 – 2025, Issue V, May 2025.
<https://www.ijraset.com/best-journal/intelligent-resume-matching-system>
- 6- N. Vanetik, Job Vacancy Ranking with Sentence Embeddings, Information, Volume 14 – 2023, Issue 8, August 2023.
<https://www.mdpi.com/2078-2489/14/8/468>
Vadim Zadykian, Bruno Andrade, Haithem Afli, Towards Explainable Job Title Matching: Leveraging Semantic Textual Relatedness and Knowledge Graphs, arxiv, 2025.
<https://www.arxiv.org/abs/2509.09522>
- 8- Hyunjin Choi, Judong Kim, Seongho Joe, Youngjune Gwon, Evaluation of BERT and ALBERT Sentence Embedding Performance on Downstream NLP Tasks, arxiv, 2021.
<https://arxiv.org/abs/2101.10642>

- 9- Priyanka Singla, Vishal Verma, An Intelligent Job Recommendation System based on Semantic Embeddings and Machine Learning, researchgate, Volume 10 – 2025 , Issue5s, December 2024.
https://www.researchgate.net/publication/388741604_An_Intelligent_Job_Recommendation_System_based_on_Semantic_Embeddings_and_Machine_Learning/link/67a45f018311ce680c563570/download?tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19
- 10- Duygu Çelik Ertuğrul, Selin Bitirim, Job recommender systems: a systematic literature review, applications, open issues, and challenges, researchgate, Volume 12 – 2025, Issue140,2025.
https://www.researchgate.net/publication/392334036_Job_recommender_systems_a_systematic_literature_review_applications_open_issues_and_challenges
- 11- Paolo Cremonesi et al, Performance of Recommender Algorithms on Top-N Recommendation Tasks, researchgate, September2010.
https://www.researchgate.net/publication/221141030_Performance_of_recommender_algorithms_on_top-N_recommendation_tasks
- 12- Ahmed Hazem Elabady, Wuzzuf Job Listings Dataset (Egypt) - January 2025.
<https://www.kaggle.com/datasets/ahmedhazemelabady/wuzzuf-job-listings-dataset-egypt-january-2025>
- 13- Gaurav Dutta, Resume Dataset2020.
<https://www.kaggle.com/datasets/gauravduttakiit/resume-dataset>
- 14- Arsh Koneru and Zoey Yu Zou, LinkedIn Job Postings (2023 - 2024).
<https://www.kaggle.com/datasets/arshkon/linkedin-job-postings>
- 15- Snehaan Bhawal, Resume dataset2021.
<https://www.kaggle.com/datasets/snehaanbhawal/resume-dataset>