Building knowledge graph for relevant degree recommendations using semantic similarity search and named entity recognition

Elkaimbillah Zineb, Mcharfi Zineb, Khoual Mohamed, El Asri Bouchra
IMS Team, ADMIR Laboratory Rabat IT Center, ENSIAS, Mohammed V University in Rabat, Rabat, Morocco

Article Info

Article history:

Received Jun 4, 2024
Revised Sep 10, 2024
Accepted Sep 29, 2024

Keywords:

BERT
Career guidance
Higher education
Knowledge graph
Named entity recognition
Recommendation
Semantic similarity search

ABSTRACT

Career guidance is a critical and often daunting process, particularly during the transition from high school to higher education within the Moroccan education system. Faced with a vast array of university programs and career options, students frequently struggle to make informed decisions that align with their aspirations and skills. To address this challenge, our research introduces an innovative system that combines semantic similarity search with knowledge graph (KG) construction to enhance the precision and personalization of academic recommendations. By utilizing Sentence-BERT (SBERT) for semantic similarity, we generate embedding vectors that capture nuanced relationships between student profiles and degree descriptions. Subsequently, named entity recognition (NER) is applied to extract essential information such as skills, fields of study, and career opportunities from these profiles and descriptions. The extracted entities and their interrelationships are then structured into a coherent KG, stored in a Neo4j database, enabling efficient querying and visualization of complex data connections. This approach provides a transparent and explainable framework, ultimately delivering tailored advice that aligns with students' individual needs and educational goals.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Elkaimbillah Zineb
IMS Team, ADMIR Laboratory Rabat IT Center, ENSIAS, Mohammed V University in Rabat
Rabat, Morocco
Email: zineb_elkaimbillah@um5.ac.ma

1. INTRODUCTION

Worldwide, academic and career guidance plays a crucial role in the personal and professional development of students [1]. It allows young people to discover their interests, identify their skills, and make informed choices regarding their future academic and professional paths. Each year in Morocco, thousands of high school graduates must choose from a multitude of university programs and career paths. The Moroccan educational system implements various initiatives to support these students in their decision-making process, including orientation sessions in schools, open house days at universities, and the use of digital platforms to provide information on the available options. As a result, this growing diversity requires advanced technologies and systems capable of providing recommendations that help students choose the options best suited to their academic and professional aspirations.

In the field of academic and career guidance, artificial intelligence (AI) provides powerful tools to help students navigate the complex choices of university paths [2]. The diversity of programs available, coupled with the variability of student profiles, creates a major challenge in aligning individual skills and aspirations with existing academic options. This challenge is amplified by the fragmentation of academic data and the lack of transparency in existing recommendation systems. The utilization of sophisticated techniques to depict and organize information becomes indispensable. Textual and unstructured data include complex knowledge [3] that provides substantial value for analysis [4]. However, retrieving information from these unstructured data is recognized as one of the least exploited opportunities in data science. It is in this context that knowledge graphs (KG) stand out as a promising approach. They offer an intuitive abstraction for representing entities and the complex relationships between them, thus facilitating a better understanding of the academic pathways available for each student.

Recommender systems using KG have already proven their effectiveness in various fields, including e-commerce, for example [5]. The study developed a personalized recommender system using an embedded KG to improve the accuracy and personalization of recommendations. This approach was found to significantly improve the relevance of recommendations and demonstrated scalability and efficiency in managing large datasets. Similarly, Loukili *et al.* [6] the authors developed a recommendation system for e-commerce using machine learning techniques. They found that these techniques improved the user experience and potentially increased online sales. In the healthcare field, Gong *et al.* [7] have developed a drug recommendation system based on the integration of a medical knowledge table, taking into account drug interactions and specific clinical contexts, reducing the risk of medical errors. In the field of education, the use of KGs has also attracted growing interest. Shi *et al.* [8], have developed a learning path recommendation model using a multidimensional knowledge table framework. This system proposes personalized pathways by integrating various dimensions, such as students' skills and interests. The authors found that this approach improves the accuracy and relevance of recommendations, offering a more tailored and effective e-learning experience. Lu *et al.* [9] presents Radarmath, an intelligent tutoring system for mathematics education. Using AI, Radarmath adapts exercises and recommendations to students' individual performance. The authors found that this system improves mathematics learning by offering personalized pathways, enabling students to progress at their own pace with targeted support for difficult concepts. Still in the research field, Zayet *et al.* [10] proposes a conceptual framework for developing personalized recommendation systems for online learning by primary and secondary school students. The authors identify current challenges, such as personalization and the integration of AI, and point out the shortcomings of existing systems. They propose solutions to overcome these obstacles, offering a guide to developers and educators in the field of e-learning. However, very little research has addressed the application of KGs specifically in the context of academic and career guidance, with one notable exception being the work of [11] proposing a recommendation system aimed at improving students' academic guidance while incorporating explanations of the recommendation. The authors have developed an approach that combines KG with collaborative filtering techniques to improve the accuracy and transparency of recommendations.

Despite the efforts deployed in the field of academic and career guidance, several gaps remain. Existing recommendation systems often suffer from a lack of fine personalization, limiting their ability to precisely align student profiles with available academic programs. Moreover, the absence of a systematic approach to integrating and structuring heterogeneous student and degree data poses a major challenge, leading to fragmented and inconsistent recommendations.

To address the challenges of guiding high school graduates toward higher education, collaborations between developers, researchers, and career guidance experts are necessary. This paper aims to develop a recommendation system to ensure better orientation and integration of high school students into higher education, particularly for scientific baccalaureate holders who require specific guidance due to the diversity and complexity of the academic and professional paths available to them. To achieve this goal, we propose an innovative solution that integrates advanced techniques of semantic similarity research and uses NER to build a coherent KG.

The main contributions of our work are as follows:

- Enhancement of recommendation personalization: We integrate semantic similarity research with SBERT to capture the contextual nuances of student profiles and degree descriptions, thus providing finer and more tailored recommendations to individual needs.
- Extraction, structuring, and transparency of recommendations: our approach uses named entity recognition (NER) to extract and structure relevant information, linking different entities with relations, thus facilitating the construction of a coherent KG, which improves the transparency and explain ability of recommendations by clearly visualizing the relationships between different entities.
- Coherent data integration: collecting and integrating specific data on Moroccan students and academic programs into the constructed KG allows for in-depth visualization and analysis of the data. This facilitates the management of fragmented data and their use for more comprehensive and consistent recommendations.

This article includes four sections. Section 2 describes the methodology used to achieve the research objective and the proposed architecture for building the recommendation system for guiding Moroccan students. Section 3 presents the results obtained and discusses the conclusions of the research. Finally, the article provides some conclusions and recommendations in section 4.

2. METHOD: DEGREE RECOMMENDATION-BASED KNOWLEDGE GRAPH CONSTRUCTION

2.1. Proposed architecture

The methodology of this research focuses on the development of a recommendation system for Moroccan students by integrating semantic similarity and KG construction techniques. It comprises five key stages as shown in Figure 1: collection of input documents, semantic similarity search, KG construction, graphical visualization of recommendations, and recommendation evaluation. This structured approach guarantees personalized, relevant guidance for students.

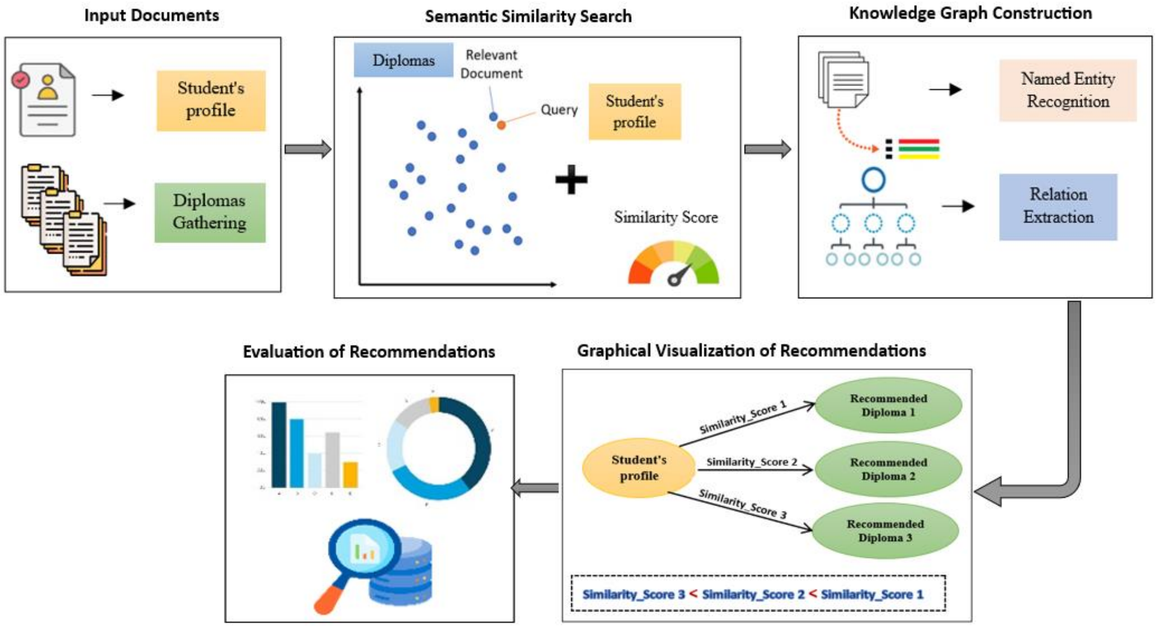


Figure 1. Proposed architecture in the KG construction process

2.2. Input data collection

Data collection is a crucial step in the development of effective academic recommendation systems. In our study, we conducted an online data collection on post-secondary degrees in Morocco. This involved researching and aggregating detailed information on various academic programs offered by higher education institutions for three levels of study (Bac+2, Bac+3, and Bac+5). The data collected includes specializations, duration of studies, employment opportunities, admission requirements, skills acquired, and modules taught.

The research utilized data from official university websites, government educational portals, and national education databases. Figure 2 visually represents the types of degrees available in the Moroccan educational system at the Bac+2, Bac+3, and Bac+5 levels, while Figure 3 displays the number of branches accumulated for each study level.

Simultaneously, we collected information on holders of a scientific baccalaureate through a survey conducted at high schools as part of their orientation process. These personal and academic data provided by the students included details such as name, baccalaureate degree obtained, academic and professional interests, skills, desired duration of studies, and career aspirations. This integrated approach to data collection enabled us to create a rich and diversified knowledge base, essential for recommending personalized educational paths to students based on their needs and aspirations.

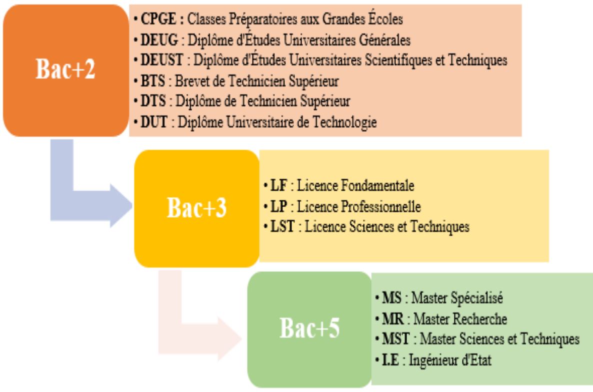


Figure 2. Type of degree used in the research sample

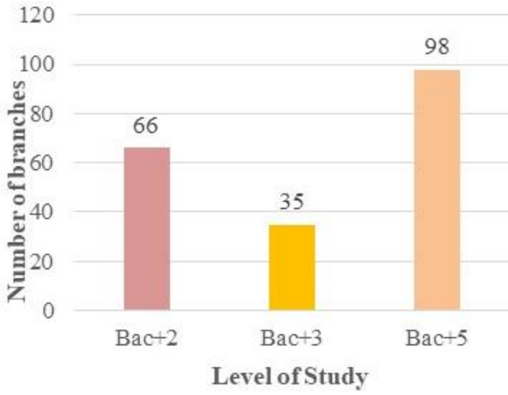


Figure 3. Number of branches per diploma level

2.3. Semantic similarity search
2.3.1. Overview and justification

Semantic similarity research [12] is a technique used in natural language processing (NLP) [13] to measure the contextual proximity between different textual units, such as words, sentences, or documents. Unlike traditional methods based on keywords, semantic similarity research considers the context and overall meaning of texts. This technique often uses deep neural network models, such as bidirectional encoder representations from transformers (BERT) and its variants like Sentence-BERT (SBERT). These models enable a more nuanced understanding of text by embedding sentences into a high-dimensional space where semantically similar sentences are located close to each other, thereby facilitating more accurate and context-aware recommendations in various applications including academic guidance systems.

BERT [14] is a deep language model pre-trained by Google, designed to understand the context of each word in a sentence by considering the words that precede and follow it. SBERT [15], an optimized variant of BERT for sentence comparison, generates effective and precise vector representations of texts. This allows for the comparison and ranking of documents based on their contextual and semantic similarity. To measure the similarity between the generated text vectors, we use cosine similarity [16], which evaluates the similarity between two vectors (a) and (b) in information retrieval by calculating the cosine of the angle between them, where each object is represented by a vector Xa and a vector Xb. This method is particularly suited for comparing high-dimensional vectors like those generated by SBERT.

$$Cos\ (Xa,Xb) = \frac{Xa.Xb}{||Xa\ ||^2 * ||Xb||^2}$$

The use of SBERT for semantic similarity enables capturing the nuanced context of student profiles and academic programs. This leads to more personalized and precise recommendations. By understanding deeper meanings in the text, this approach enhances the effectiveness of academic guidance systems, rather than relying solely on keyword matching.

2.3.2. Implementing the semantic similarity method

In this section, we describe in detail the steps involved in implementing the semantic similarity method used to align student profiles with academic program descriptions. Figure 4 shows the process of using the semantic similarity solution with SBERT. Each step is crucial to ensure the accuracy and efficiency of the recommendation system.

A. Document parser

Document parsing is the initial step in our process. This stage involves processing academic diplomas and student profiles, converting them into a standardized format. This ensures the documents are ready for use in subsequent stages.

B. Preprocessing

Data pre-processing involves several sub-steps aimed at cleaning and normalizing texts for better analysis. This includes the removal of special characters, tokenization, lemmatization and the elimination of stop words. This process ensures that text data is consistent and ready for feature extraction and semantic comparison.

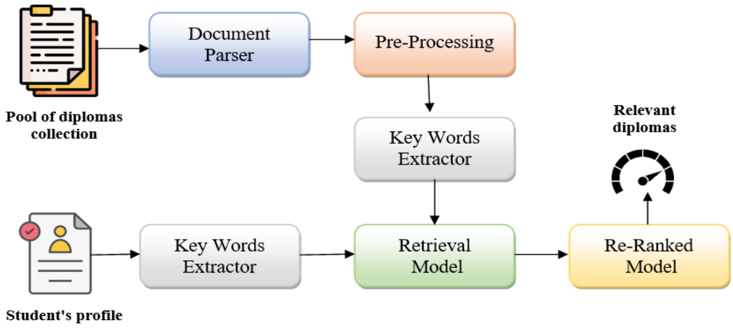


Figure 4. Procedure for implementing the semantic similarity solution

C. Keywords extractors

Keyword extraction is performed using the term frequency-inverse document frequency (TF-IDF) method. TF-IDF [17] is a statistical measure used to evaluate the importance of a word in a document in relation to a corpus. It combines the frequency of appearance of a term in a document (TF) and the inverse frequency of the document in the corpus (IDF), enabling words to be weighted according to their importance. This technique enables us to identify the most characteristic and significant terms in student profiles and academic program descriptions, thus facilitating the following semantic similarity steps.

D. Retrieval model

The retrieval model is the first step in the recommendation process, where the aim is to retrieve an initial list of documents (academic programs) that are relevant to a query (student profile). Cosine similarity is often used at this stage to measure the similarity between the query vector and the document vectors in the database.

- Data encoding: student profiles and degree descriptions are transformed into numerical vectors using SBERT models.
- Cosine similarity calculation: for each student profile, the cosine similarity is calculated between the profile vector and the vectors of all available diplomas.
- Document retrieval: the diplomas are then sorted in descending order of cosine similarity, and the most similar are selected to form the initial list of recommendations.

E. Re-ranked model

The re-rank model is a subsequent step in which the initial list of retrieved documents is refined and re-ranked to improve the accuracy of the recommendations. This step again uses cosine similarity, but also takes into account other factors such as explicit student preferences and assigns a relevance score to each document (academic program)-query (student profile) pair. These scores indicate the degree of relevance of the query to the document and are used to reorganize the initial search results in the reclassified model step. The end result is an optimized list of recommendations customized for each student.

2.4. Knowledge graph construction

The construction of KG, inspired by human intelligence and problem-solving methods, offers a powerful structured representation of facts, involving the creation of a structured representation of knowledge by identifying entities and their relationships within a data set [18]. It is particularly useful in the field of education [19]. This approach uses predefined entity labels and organizes data in the form of a graph, facilitating complex queries and data analysis. By structuring and integrating complex information about students and academic programs, KG offer a unified framework for analysis and a comprehensible rationale for recommendations.

2.4.1. Conceptualization of the data model

In this section, we describe the conceptual model for building a KG for an academic recommendation system. This model visualizes relationships between different entities such as student profiles and degree descriptions, as well as associated skills and job opportunities. It encompasses eight main entities: student profile, skills/interests, bachelor's degree, career aspirations, degree description, branch, acquired skills and job opportunities, each with its own attributes. The data model is represented visually in Figure 5 and detailed in Table 1.

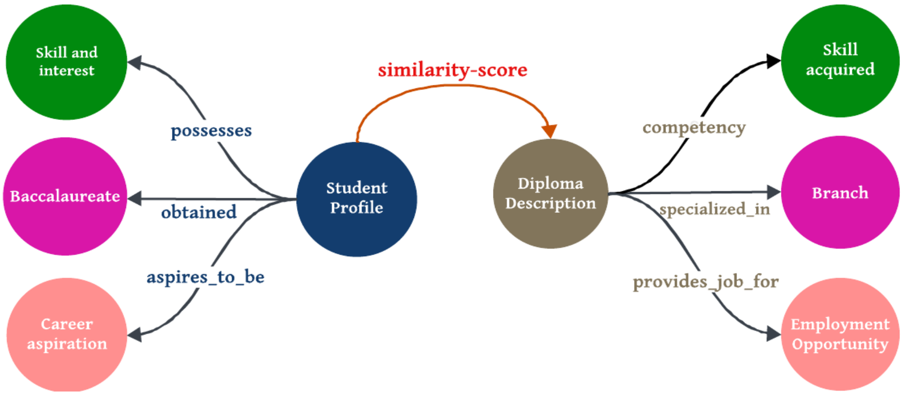


Figure 5. Conceptual graphical data model

Table 1. Description of a conceptual graphical data model

Entity	Entity description	Property	Property description
Student profile	represent key information about student profiles	id Student name Desired study duration	Unique identifier of the student profile Student name: name of the student Represent the duration of study that the student desires to pursue after their baccalaureate.
Baccalaureate	Represents key information regarding the type and specialization of the baccalaureate qualification obtained by the student	Type of baccalaureate Branch of the baccalaureate Mention obtained	The type of Baccalaureate. The specialization of the baccalaureate. The mention obtained in the baccalaureate
Career aspiration	This attribute will contain information about the career aspirations or professional goals of the student	Career aspiration description	Contains information about the career aspirations or professional goals of the student.
Skill/Interest	List of skills or areas of expertise and academic or professional interests of the student	Skill Interest	List of skills or areas of expertise of the student Academic or professional interests of the student
Diploma description	A diploma description provides a comprehensive overview of a diploma program, detailing the essential aspects that define the educational experience and outcomes	id Diploma name Admission requirement Study duration	Unique identifier of the diploma Official name of the diploma Admission criteria for this diploma Represent the duration of the diploma program in years
Branch	The specialization of the diploma	Specialization Module	The specialization of the diploma. This attribute will contain information about the modules or courses offered as part of the diploma program.
Employment opportunity	Professional opportunities associated with this diploma	Job title Job description	The title of the job associated with this diploma Description of the job role and responsibilities
Skills acquired	Provides detailed information about the skills gained through completing the diploma program	Skill name	Name of the skill acquired

2.4.2. NER and relationship extraction

NER [20] is an automatic NLP technique for automatically identifying and classifying specific entities in text. In our academic recommendation system, NER utilizes the BERT model to accurately identify and annotate key entities such as degrees, admission requirements, and career aspirations, effectively capturing contextual information for precise entity categorization. Figure 6 illustrates BERT's underlying architecture specifically designed for the NER task, highlighting the different layers and mechanisms of the model.

Once the named entities have been identified, the next step is to extract the semantic relationships between them. This enables us to understand how entities interact with each other in the context of degree descriptions and student profiles. The semantic relations extracted are then structured in the form of triplets (subject, relation, object) to facilitate their integration into the KG. Among these relations, the similarity-Score relation is particularly important. This relationship, obtained from the semantic similarity search method, measures the contextual proximity between the student profile and the diploma description. The similarity-score relationship quantifies how relevant a diploma is to a given student, based on an in-depth semantic analysis of the texts. Identifying the relationships between entities enables us to understand the interactions and dependencies between them, which is essential for building rich, informative KG.

The construction of KG plays a crucial role in improving the transparency and explicability of recommendations. By structuring information into nodes and relations, our recommendation system offers explicable and justified suggestions. This enables users to understand why certain recommendations are made, thereby increasing confidence in the system.

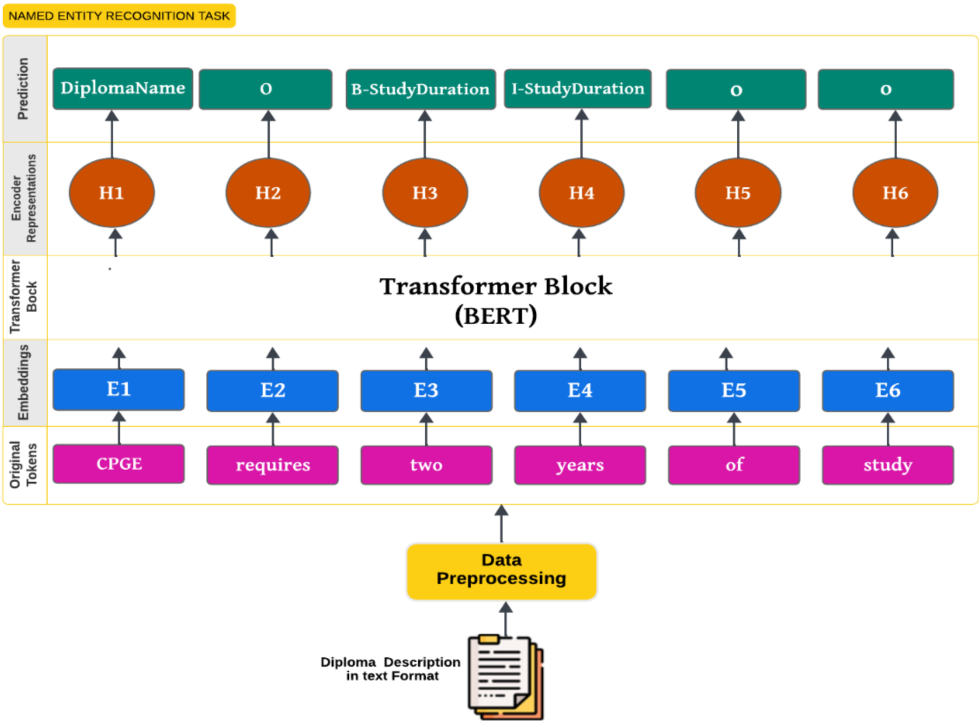


Figure 6. Architecture of BERT for NER task

2.4.3. Performance evaluation

The evaluation and validation of our academic recommendation system are crucial steps in guaranteeing efficiency and accuracy and user satisfaction. This section describes the methodologies used to evaluate the system's performance and validate the recommendations provided.

A. Quantitative evaluation

To evaluate the performance of our NER model, we used accuracy as the main measure. Accuracy measures the ratio between the number of correct positive predictions and the total number of positive predictions, providing a clear assessment of the model's accuracy in identifying relevant entities. This measure is particularly useful for minimizing false positives. It enables us to quantify the effectiveness of the NER model in identifying and classifying entities from descriptions of university programs and student profiles. It is calculated as follows:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

B. Qualitative evaluation

In addition to the quantitative evaluation, we conducted a qualitative evaluation to measure student satisfaction with the recommendations provided by our system. This qualitative evaluation was carried out through questionnaires distributed to students after they had used the system. Key satisfaction indicators included the relevance of the recommendations, the clarity of the explanations provided by the system, and ease of use.

3. RESULTS AND DISCUSSION

3.1. Libraries and technical environment

We developed our algorithm using Python, a key language in the field of AI. Python, a popular AI language for its ease of installation and speed. Python's intuitive syntax allows for rapid learning and shorter development cycles. The sentence-transformers library was used for semantic similarity search, generating sentence embeddings using pre-trained models like BERT. The study utilized Hugging Face's transformers library [21] for NER with BERT, pandas for structured data manipulation, numpy for numerical calculations, keras for neural network development, and sklearn for machine learning tasks like performance evaluation and data pre-processing. These libraries provide user-friendly interfaces for transformer-based models and support multidimensional arrays and complex mathematical operations. The environment in which the models were trained and run was a Windows 11 Professional version 22H2, equipped with a Core i7 processor, 16 GB DDR4 RAM and two graphics cards, including NVIDIA GeForce MX350 and INTEL(R) Iris Xe Graphics.

To put the idea into practice, the graphics database was created using Neo4j, a powerful tool for handling complex relationships and graph traversal. Cypher [22], a query language specifically designed for graphical databases, was used for data analysis. Neo4j's efficient handling of graph relationships and high performance make it ideal for visualization and manipulation of structured data. Cypher allows users to define templates for efficient query and mutation operations.

3.2. Results: case study

3.2.1. Semantic similarity result

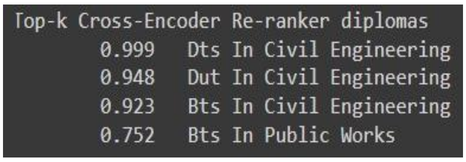
Our approach is to apply a semantic similarity model, based on SBERT technology, to a dataset comprising both undergraduate student profiles and a comprehensive database of post-baccalaureate degrees available. We use this model to calculate similarity scores between each student profile and the degrees listed. These similarity scores provide a quantitative measure of the relevance of each diploma to a specific student profile.

In our results, we present a list of the most relevant degrees recommended for each student profile. This list is ranked in descending order according to the similarity scores associated with each recommendation. These recommendations were generated on the basis of several factors in the student profile, such as desired length of study, skills and interests, career aspirations and degree information, such as branch and honors obtained in relation to available university programs. Take an example of a profile of a high school student shown in Figure 7. From this profile, we have applied our semantic similarity research method with SBERT to recommend the most relevant academic degrees among those available in Morocco. The degrees shown in Figure 8 were recommended for this student, accom panied by their similarity scores.

The results show that the DTS in Civil Engineering is the most relevant degree for the student, with a similarity score of 0.99, providing skills in design, site management, communication, and team management. The DUT in Civil Engineering offers practical site management and the use of computer-aided design software use, while the BTS in Civil Engineering has a slightly lower score, offers in-depth training in complex structural modeling and simulation. The BTS in Public Works is recommended due to its focus on public works and infrastructure, but is slightly less relevant due to its specific focus.

```
Student profile:
-  Id: 120
-  Interests/skills:
   Mathematics, Physics,Technical Problem Solving,
   Construction and Design,Teamwork,drawing,English,
   calculation,building,creation,management,supervising,
   Reading technical plans,teamwork skills,communication
-  Professional aspirationsfor Job Opportunity  :
   Work in the civil engineering sector,construction technician,
   project manager,
   supervisor,
   building designer
-  Type of Baccalaureate:  scientific baccalaureate
-  Baccalaureate specialization: physics chemistry Sciences
-  Grade obtained: Good
-  Desired length of study:2 years
```

Figure 7. Sample profile of a student with a high school diploma



```
Top-k Cross-Encoder Re-ranker diplomas
        0.999   Dts In Civil Engineering
        0.948   Dut In Civil Engineering
        0.923   Bts In Civil Engineering
        0.752   Bts In Public Works
```

Figure 8. Results of similarity scores on the search sample

3.2.2. NER performance

Table 2 shows the performance of the BERT model for the designated entities recognition task (NER). Metrics include validation loss, training loss average, and accuracy. The BERT model achieved a validation loss of 0.039, an average training loss of 0,02, an accuracy of 90%.

Table 2. Results of performances for NER

Task	Model	Validation loss	Average train loss	Accuracy
NER	BERT	0,039	0,02	0,90

These results indicate that the BERT model was effective for the NER task, with high accuracy, demonstrating the model's ability to correctly identify and classify entities named in the text. The small loss of validation and training shows that the model is well trained and does not suffer from overlearning. This is important for the quality of the data used in building the knowledge chart.

3.2.3. Graph building

We use Neo4j to create a graph database, which we then review using graph analysis methods and the Cypher query language. This sub graphic excerpt from Figure 9 illustrates the relationships between different essential nodes of a student's educational and professional journey. The recommended degrees for the student profile are highlighted in blue, while the student profiles are highlights in yellow. By integrating these elements, the subgraph becomes a powerful decision-making tool, enabling you to visualize and understand the complex relationships between a student's academic career, skills acquired, and potential career opportunities. It offers a holistic perspective to align students' aspirations and capabilities with academic offerings.

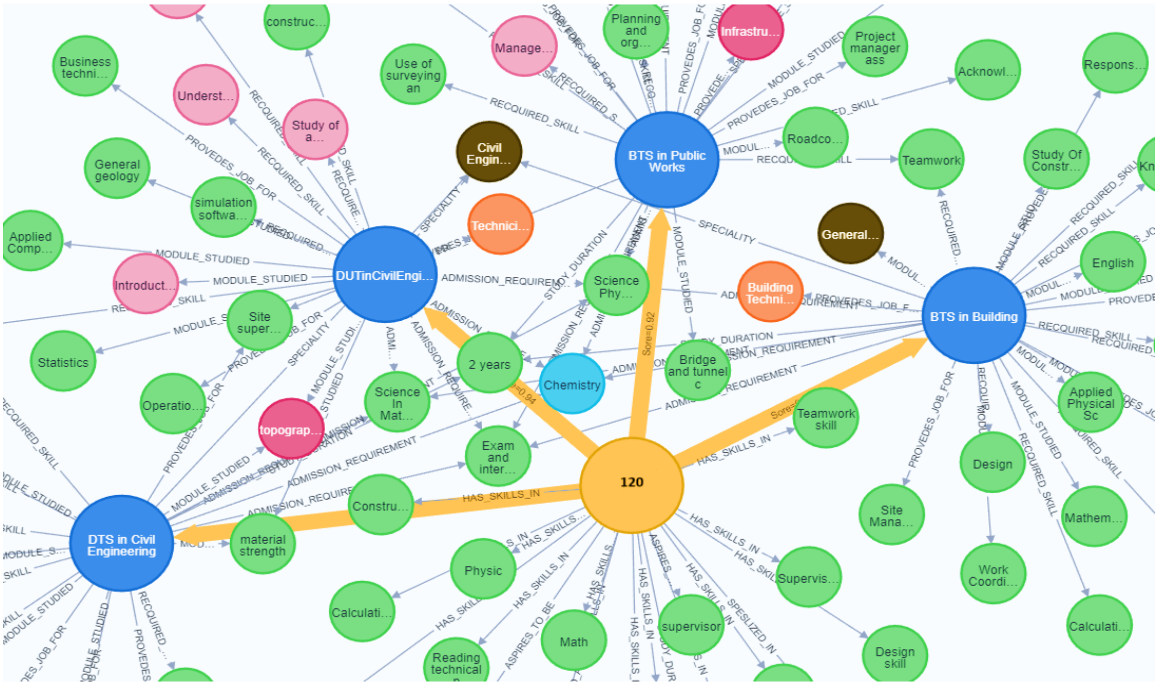


Figure 9. Excerpt from sub-graph

3.3. Qualitative evaluation of recommendations

To assess the quality of the academic degree recommendations provided by our system, we conducted a satisfaction survey among students who recently used it. The survey evaluated user satisfaction based on four criteria: how well the recommendations matched the students' skills, interests, and aspirations; the clarity and explain ability of the recommendations; the perceived impact on their future careers; and the reliability of the information provided. The results of this survey were visualized in the form of a histogram as shown in Figure 10, providing an overview of the average scores for each criterion. It illustrates the overall positive performance of the system, with strong results in terms of clarity of recommendations and suitability to the profile of students.

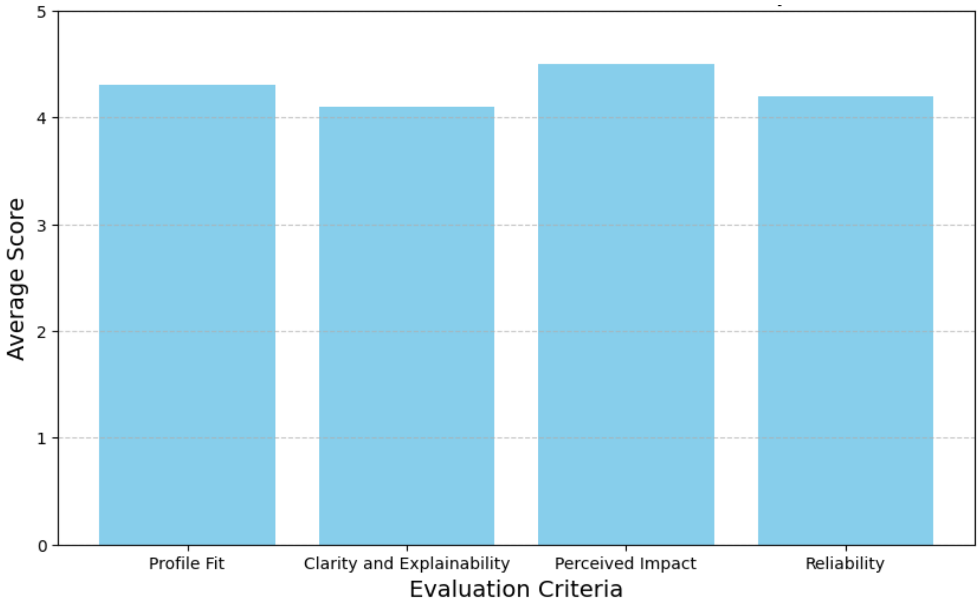


Figure 10. Qualitative evaluation of the academic recommendation system

3.4. Model comparison and limitation

In our study, we evaluate our technique, which relies on semantic similarity and KG, by comparing it to various methodologies employed in academic orientation recommendation systems. This comparison examines variables such as customization, clarity, data incorporation, the considerations taken into account, and the specific sample being targeted as shown in Table 3.

Table 3. Comparative study of methods used in academic orientation

Ref / criteria	Method	Personalization	Transparency	Data integration	Targeted sample
[23]	Machine learning models	High, based on ML predictions	Low, complex models with limited transparency	Low, focused on predictive modeling without deep integration	Finland student population
[24]	Fuzzy intelligence	Medium, based on academic features	Medium, recommendations based on rules and algorithms	Partial, limited to academic data	Palestine Engineering students
[25]	SVM model	Medium, based on input features	Low, complex model difficult to explain	Low, mainly academic and behavioral data	University students
[26]	XGBoost Model	High, dependent on academic performance	Medium, requires advanced interpretations	Medium, integrates academic performance	University students
Our approach	Similarity semantic Ner + KG	Very high, captures contextual nuances via semantic similarity search	High, through the explainability of relationships via KG	Coherent integration of complex data via KG	Scientific Baccalaureate students in Morocco

The table clearly shows that our approach stands out from other methods primarily due to its ability to offer high personalization, coherent data integration, and enhanced transparency through the use of KG. While the other methods, though effective in certain contexts, suffer from limitations in terms of fine personalization, integration of fragmented data, and the explain ability of recommendations. These elements position our approach as an innovative and more comprehensive solution for academic orientation, particularly in the context of Moroccan scientific baccalaureate students.

Although our study demonstrated promising results, some limitations remain. For example, the BERT-based NER model, while effective, may exhibit biases in entity recognition due to linguistic specificities or variations in qualification descriptions. Furthermore, the quality of recommendations is highly dependent on the input data, and gaps in available data could limit the system's ability to generalize its recommendations. To address these challenges, it would be beneficial to extend the research to a broader range of variables, such as labor market trends, economic forecasts, and geographic constraints. This approach would allow for further refinement of recommendations and ensure a better fit between students and their future academic and professional paths.

4. CONCLUSION

This study presents a novel recommendation system for science student counseling, integrating semantic similarity, NER, and KG techniques. This approach overcomes the limitations of existing systems by providing personalized, transparent, and understandable recommendations. The research highlights the importance of personalizing academic pathways, demonstrating that understanding student profiles and complex relationships between programs can significantly improve student counseling. The study proposes a substantial improvement in the quality of recommendations and opens new research avenues, such as the integration of advanced AI models for even more refined personalization. The results have important implications for the field of school counseling, as they show that NLP techniques and knowledge graphs can transform guidance systems, making them more efficient and tailored to individual needs.

REFERENCES

[1] P. Roy, "Career guidance: a way of life," *SSRN Electronic Journal*, 2020, doi: 10.2139/ssrn.3640339.
[2] S. Purnomo and R. Gunaningrat, "Determinants of student interest in choosing a study program," *International Journal of Social Science*, vol. 1, no. 6, pp. 873–878, Apr. 2022, doi: 10.53625/ijss.v1i6.1899.
[3] J. Fan, A. Kalyanpur, D. C. Gondek, and D. A. Ferrucci, "Automatic knowledge extraction from documents," *IBM Journal of Research and Development*, vol. 56, no. 3–4, pp. 5:1-5:10, May 2012, doi: 10.1147/JRD.2012.2186519.
[4] R. J. Mooney and R. Bunescu, "Mining knowledge from text using information extraction," *ACM SIGKDD Explorations Newsletter*, vol. 7, no. 1, pp. 3–10, Jun. 2005, doi: 10.1145/1089815.1089817.
[5] N. L. Le, M. H. Abel, and P. Gouspillou, "A personalized recommender system based-on knowledge graph embeddings," in *Lecture Notes on Data Engineering and Communications Technologies*, vol. 164, Springer Nature Switzerland, 2023, pp. 368–378.
[6] M. Loukili, F. Messaoudi, and M. El Ghazi, "Machine learning based recommender system for e-commerce," *IAES International Journal of Artificial Intelligence*, vol. 12, no. 4, pp. 1803–1811, Dec. 2023, doi: 10.11591/ijai.v12.i4.pp1803-1811.
[7] F. Gong, M. Wang, H. Wang, S. Wang, and M. Liu, "SMR: medical knowledge graph embedding for safe medicine recommendation," *Big Data Research*, vol. 23, p. 100174, Feb. 2021, doi: 10.1016/j.bdr.2020.100174.
[8] D. Shi, T. Wang, H. Xing, and H. Xu, "A learning path recommendation model based on a multidimensional knowledge graph framework for e-learning," *Knowledge-Based Systems*, vol. 195, p. 105618, May 2020, doi: 10.1016/j.knosys.2020.105618.
[9] Y. Lu, Y. Pian, P. Chen, Q. Meng, and Y. Cao, "RadarMath: an intelligent tutoring system for math education," *35th AAAI Conference on Artificial Intelligence, AAAI 2021*, vol. 18, no. 18, pp. 16087–16090, May 2021, doi: 10.1609/aaai.v35i18.18020.
[10] T. M. A. Zayet, M. A. Ismail, S. H. S. Almadi, J. M. H. Zawia, and A. M. Nor, "What is needed to build a personalized recommender system for K-12 students' E-learning? recommendations for future systems and a conceptual framework," *Education and Information Technologies*, vol. 28, no. 6, pp. 7487–7508, Dec. 2023, doi: 10.1007/s10639-022-11489-4.
[11] N. Hubert, A. Brun, and D. Monticolo, "Vers un système de recommandation explicable pour l'orientation scolaire," *Workshop EXPLAIN'AI-EGC Blois 2022*, 2022.
[12] A. De Nicola, A. Formica, M. Missikoff, E. Pourabbas, and F. Taglino, "A comparative assessment of ontology weighting methods in semantic similarity search," in *ICAART 2019 - Proceedings of the 11th International Conference on Agents and Artificial Intelligence*, 2019, vol. 2, pp. 506–513, doi: 10.5220/0007342805060513.
[13] T. Wolf *et al.*, "Transformers: state-of-the-art natural language processing," in *EMNLP 2020 - Conference on Empirical Methods in Natural Language Processing, Proceedings of Systems Demonstrations*, 2020, pp. 38–45, doi: 10.18653/v1/2020.emnlp-demos.6.
[14] J. Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
[15] C. Hu, X. Sun, H. Dai, H. Zhang, and H. Liu, "Research on log anomaly detection based on sentence-BERT," *Electronics (Switzerland)*, vol. 12, no. 17, p. 3580, Aug. 2023, doi: 10.3390/electronics12173580.
[16] K. Bagheri Fard, M. Nilashi, and N. Salim, "Recommender system based on semantic similarity," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 3, no. 6, Dec. 2013, doi: 10.11591/ijece.v3i6.3931.
[17] P. Bafna, D. Pramod, and A. Vaidya, "Document clustering: TF-IDF approach," in *International Conference on Electrical, Electronics, and Optimization Techniques, ICEEOT 2016*, Mar. 2016, pp. 61–66, doi: 10.1109/ICEEOT.2016.7754750.

[18] S. Ji, S. Pan, E. Cambria, P. Marttinen, and P. S. Yu, "A survey on knowledge graphs: representation, acquisition, and applications," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 2, pp. 494–514, Feb. 2022, doi: 10.1109/TNNLS.2021.3070843.

[19] Z. Elkaimbillah, M. Rhanoui, M. Mikram, and B. El Asri, "Comparative study of knowledge graph models in education domain," in *Proceedings of the 2nd International Conference on Big Data, Modelling and Machine Learning*, 2022, pp. 339–344, doi: 10.5220/0010733800003101.

[20] Z. Nasar, S. W. Jaffry, and M. K. Malik, "Named entity recognition and relation extraction: state-of-the-art," *ACM Computing Surveys*, vol. 54, no. 1, pp. 1–39, Feb. 2021, doi: 10.1145/3445965.

[21] T. Wolf, "Huggingface's transformers: State-of-the-art natural language processing," *arXiv preprint arXiv:1910.03771*, 2019.

[22] J. Guia, V. G. Soares, and J. Bernardino, "Graph databases: Neo4j analysis," in *ICEIS 2017 - Proceedings of the 19th International Conference on Enterprise Information Systems*, 2017, vol. 1, pp. 351–356, doi: 10.5220/0006356003510356.

[23] A. Dirin and C. A. Saballe, "Machine learning models to predict students' study path selection," *International Journal of Interactive Mobile Technologies*, vol. 16, no. 1, pp. 158–183, Jan. 2022, doi: 10.3991/IJIM.V16I01.20121.

[24] M. Qamhieh, H. Sammaneh, and M. N. Demaidi, "PCRS: personalized career-path recommender system for engineering students," *IEEE Access*, vol. 8, pp. 214039–214049, 2020, doi: 10.1109/ACCESS.2020.3040338.

[25] H. Q. Nguyen, D. D. K. Nguyen, T. D. Le, A. Mai, and K. T. Huynh, "Career path prediction using XGBoost model and students' academic results," *CTU Journal of Innovation and Sustainable Development*, vol. 15, no. ISDS, pp. 62–75, Oct. 2023, doi: 10.22144/ctujoisd.2023.036.

[26] Z. Wang, G. Liang, and H. Chen, "Tool for predicting college student career decisions: an enhanced support vector machine framework," *Applied Sciences (Switzerland)*, vol. 12, no. 9, p. 4776, May 2022, doi: 10.3390/app12094776.

BIOGRAPHIES OF AUTHORS

Elkaimbillah Zineb received a degree in Computer Sciences (ENSIAS) in 2019. She is currently a Ph.D. student in the IMS (IT Architecture and Model Driven Systems Development team) Team of Advanced Digital Enterprise Modeling and Information Retrieval Research Laboratory (ADMIR Laboratory) at ENSIAS. Her research interests include semantic web, knowledge graph, ontology, information representation and machine learning/deep learning approach in education domain. She can be contacted at email: zineb_elkaimbillah@um5.ac.ma.

Prof. Dr. Mcharfi Zineb is a Computer Science professor and member of the IMS Team, ADMIR Laboratory Rabat IT Center at ENSIAS, Mohammed V University in Rabat; Morocco. She has directed and continues to direct numerous doctoral theses in several themes. Her research interests include the application of artificial intelligence in education, software product line engineering, agile software development and software traceability. She can be contacted at email: zineb_mcharfi@um5.ac.ma.

Khoual Mohamed received a degree in networks and computer systems from Faculty of Sciences and technology Settat (FST) in 2018. He is currently a Ph.D. student in the IMS Team of ADMIR Laboratory at ENSIAS. His research interests include machine learning approach in education domain and data mining. He can be contacted at email: mohamed_khoual@um5.ac.ma.

Prof. Dr. El Asri Bouchra is currently director of teaching-research at ENSIAS. She holds various positions, such as head of the software engineering department and coordinator of the Software Engineering program at ENSIAS, she has supervised and continues to supervise numerous doctoral theses in software architecture and data management in the health, industrial, and education sectors. She is a Professor in the Software Engineering Department and a member of the IMS Team at ENSIAS. Her expertise and contributions extend to scientific committees, doctoral study centers, and teaching modules of the ENSIAS Software Engineering program. Her research interests include service-oriented computing, model driven engineering, cloud computing, component-based systems and software product line engineering. She can be contacted at email: bouchra.elasri@ensias.um5.ac.ma or b.elasri@um5s.net.ma.