AI-Powered Resume Intelligence System Using LLM and Retrival Augmented Generation

Kamalaveni V[1], Midunavarsini.B[2], Sarayuma .M [3], Shikha Srinivas[4], Sooriya .G.M[5]

1*Assistant Professor, Department Of Artificial Intelligence and Data Science, Sri Shakthi Institute of Engineering and Technology, Coimbatore, Tamil Nadu, India.

2,3,4. Fourth Year B-Tech AI&DS, Sri Shakthi Institute of Engineering and Technology, Coimbatore, Tamil Nadu, India.

ABSTRACT

The fast-changing nature of the labor market is driving the need to have intelligent systems that are able to effectively determine the skill profiles of individuals and match the ever-changing needs in the industry. The conventional Applicant Tracking Systems (ATS) are based on keyword-based matching, which frequently does not reflect the semantic associations between the skills acquired by candidates and the job description. The current paper presents a proposal for an Adaptive Career Intelligence Agent that combines Multi-Source Retrieval-Augmented Generation (RAG) with Explainable Artificial Intelligence (XAI) to provide context-relevant and trustworthy career data. It uses a dual-source retrieval system that processes user resumes and industry job descriptions in a FAISS-based vector database with HuggingFace semantic representations and similarity search. A big language model (Gemini 1.5 Flash) does structure skill classification, classifying the skills into Core, Competitive and Future. An explainability layer makes all outputs have a foundation on retrieved sources, enhancing transparency and minimizing hallucinations. Also, there is a career trajectory module that determines the skill gaps and suggests individual development routes. The proposed architecture is scalable, cost-effective and applicable to the real-world implementation.

KEYWORDS

FastAPI, natural language processing, python, RAG

INTRODUCTION

In The growing sophistication in the current workforce has posed an immense disconnection between personal skills and constantly changing industry demands. The existing traditional recruitment and career advice systems heavily depend on methods of matching based on words and photos, and these methods are not always effective in identifying the semantic associations between job descriptions and the profiles of applicants. This leads to poor skills evaluation and lost career prospects.

The high rate of technological progress has also increased the necessity to have smart systems that can help individuals match their competencies to the industry's needs. Natural Language Processing (NLP) and Artificial Intelligence (AI) are important tools to overcome these issues since it can analyze unstructured text data (resumes and job descriptions). Machine learning methods are used to recognize trends and connections between proficiencies and work demands and enhance the precision of suggestions. Also, semantic search and vector-based representations supplement contextual

Understanding to a type of simple keyword matching. To address such constraints, this paper suggests the use of an Adaptive Career Intelligence Agent built on a Multi-Source Retrieval-Augmented Generation (RAG) model. The system simultaneously compares resumes with job descriptions with the help of semantic embeddings and categorizes skills as core, competitive, and future. Also, the inclusion of Explainable Artificial Intelligence (XAI) is used to ensure transparency to give source-grounded justifications. There is also a module of career trajectory prediction which provides individual growth path and thus filling the gap between personal capacity and the changing market expectations.

LITERATURE REVIEW

"Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks" by Lewis et al. (2020)

Retrieval-Augmented Generation (RAG) by Lewis et al. (2020) was presented as a viable remedy to one of the major shortcomings of conventional language models. Most traditional models will not look beyond what they were taught in their training, and this implies that their responses may become outdated or inaccurate at times. To counter this, RAG is a hybrid of two potent concepts information retrieval and text generation such that the model can look up information that is relevant and then generate an answer. In this method, the system will first find helpful documents through semantic similarity that is, the system perceives the context of the query and not the matching of the keywords. It then applies this information retrieved as a reference in generating a response. This renders the output more relevant, accurate and reliable. It was experimentally demonstrated that RAG outperforms the standard transformer models, particularly in such tasks as open-domain question answering. The concept can be of great benefit in areas like resume analysis where context and accuracy of insight are extremely essential. RAG is a powerful basis on which to develop intelligent and reliable solutions by integrating retrieval and generation, which minimize errors and enhances confidence in AI systems.

"Sentence-BERT: Sentence Embeddings using Siamese BERT Networks" by Reimers and Gurevych (2019)

Reimers and Gurevych (2019) introduced Sentence-BERT as an improved version of the original BERT model, specifically designed to make sentence comparison faster and more efficient. In traditional BERT, comparing two sentences requires running the model multiple times, which can be slow and computationally expensive. Sentence-BERT solves this problem by using a Siamese network structure, allowing it to process sentences independently and then compare them directly. The model converts sentences into dense vector representations that capture their meaning in a numerical form. These vectors make it easy to measure how similar two pieces of text are using techniques like cosine similarity. This approach not only speeds up the process but also maintains high accuracy in understanding context and meaning. The authors tested Sentence-BERT on several benchmark datasets and found that it performs exceptionally well in tasks such as semantic search, clustering, and sentence similarity. Its ability to understand the meaning behind text rather than just matching words makes it highly useful in real-world applications. In systems like resume-job matching, Sentence-BERT plays a key role by accurately identifying how closely a candidate's skills align with job requirements, even when different wording is used.

"Dense Passage Retrieval for Open-Domain Question Answering" by Karpukhin et al. (2020)

Karpukhin et al. (2020) presented a more sophisticated method of locating relevant information than the classic keyword-based methods in the form of Dense Passage Retrieval (DPR). DPR pays attention to meaning of a text rather than just

matching words like the older systems like BM25. It achieves such by transforming the query of the user as well as the documents into dense numerical vectors, enabling the system to compare them to each other based on semantic similarity and not on the exact words. The dual-encoder design is one of the main characteristics of DPR, with queries and documents being represented in different but equivalent vectors. The design allows management of large-scaled data, and at the same time, it has a high level of performance. The model is fast to recall the most applicable passages with different wording (but similar meaning). Results obtained through experiments demonstrated that DPR is significantly more effective than conventional retrieval strategies in the accuracy and relevance. This is particularly handy in the case of applications that are based on semantic search. DPR is also useful in the analysis of resumes systems to effectively match user profiles with job descriptions by gaining insight into the context of skills and requirements. Generally, the research demonstrates that embedding based retrieval techniques can enhances access to information and aid in decision making.

"Attention is All You Need" by Vaswani et al. (2017)

Transformer architecture by Vaswani et al. (2017) has caused a significant change in the design of natural language processing systems. The only models that had been used previously were the recurrent or convolutional networks, which took the text character at a time and could not always handle long sequences. Transformer transformed this by introducing a mechanism known as self-attention, which enables the model to look at the entire words of a sentence simultaneously, and comprehend their relationship with each other. This method is much quicker and more efficient in processing and helps the model to be much more context and meaning sensitive. Due to these benefits, the Transformer is now at the core of numerous language models in modern times such as BERT, GPT or Gemini. Its advantage is that it can deal with long-range dependencies and has a better sense of context, which is necessary to solve such tasks as text classification, summarization, and semantic analysis. Transformer-based models are also useful in finding the correct skills and understanding their applicability in each context in applications like resume intelligence systems. This will eventually result in accurate analysis and improved system performance. The experimental findings indicated that DPR is more effective than the conventional retrieval techniques in terms of accuracy and relevance. This is particularly handy with applications that depend on semantic search. DPR can be used in the resume analysis system to better match the user profile to the job description based on the underlying context of skills and requirements. Altogether, the research demonstrates that embedding-based retrieval techniques may be highly effective in facilitating access to information and making more effective decisions.

"Language Models are Few-Shot Learners" by Brown et al. (2020)

Brown et al. (2020) introduced GPT-3, a large-scale language model that could do many tasks related to natural language with the least amount of task-specific training. The experiment has shown that huge language models are able to generalize between tasks through few-shot learning and do not require many labelled datasets. This is because the model can produce coherent and context-aware responses and is therefore applicable in text analysis, recommendation systems, and conversational AI. These models may be employed in the analysis of user profiles, skills classification, and the production of individual career guidance in resume intelligence systems. The study demonstrates the possibilities of large language models to develop intelligent and adaptive systems.

SOFTWARE COMPONENTS

1. Python
Python provides a versatile environment for data manipulation and cleaning. It supports a wide range of libraries, ensuring a robust and efficient preprocessing workflow.

2. Pandas
Pandas are used to load and preprocess the dataset, handling missing values, transforming data types, and organizing the data for model training. Its powerful Data Frame structure simplifies complex operations.

3. NumPy
NumPy is essential for numerical operations like feature scaling and array manipulation. It supports efficient computation, particularly for large datasets requiring mathematical transformations.

4. FastAPI
The system has high-performance APIs created with FastAPI. It also allows effective interaction among various modules, permits asynchronous processing and maintains quick reaction time. FastAPI is specifically appropriate to deploy scalable AI services.

5. Streamlit
Streamlit offers a user interface that allows interactive frontend, enabling the user to upload resumes and visualize the results. It makes it easy to develop easy to use dashboards and allows real-time communication with the system outputs.

6. HuggingFace Transformers
It is a library that contains pre-trained transformer models to be applied in text processing and embedding generation. It allows state-of-the-art natural language processing and integrates with massive language models to make intelligent analysis.

7. Sentence-Transformers
Sentence-Transformers is a system that produces semantic representations of text data. These embeddings learn contextual relationship between words and phrases and therefore allow similarity-based retrieval and comparison.

8. FAISS

FAISS (Facebook AI Similarity Search) is a vector database that is an efficient way to store and retrieve embeddings. It helps to quickly search for similarities on a large dataset, which is key to executing the RAG pipeline.

SYSTEM FLOW

1. Data Collection

The initial one is to gather user resumes and job descriptions. Users post their resumes and various websites acquire job descriptions where job postings can be found on online employment sites and company websites.

- **Resume Data:** User-uploaded documents in PDF or DOCX format
- **Job Descriptions:** Industry-specific requirements and skill listings
- **Mixed Data:** Combination of resume and job datasets for comparison.

2. Preprocessing
Preprocessing is essential in preparing textual data to be effectively analyzed. This is important to ensure that data is clean and organized because resumes and job descriptions usually have noise, inconsistent formatting, and irrelevant information. The first step is the text cleaning to eliminate undesired symbols, special characters, and redundant material. It is then followed by normalization where the text is standardized so that it is consistent throughout the dataset. The text is then broken down into smaller meaningful units or words or phrases through tokenization. Lastly, the removal of stop-words gets rid of words that are frequently used, but not of much value to the analysis, thus enhancing processing speed and enhancing model performance.

3. Feature Extraction
In feature extraction, the concern is to transform textual information into meaningful representations,

which is machine learning processable. Sentence-Transformers are employed in this system to produce semantic meaning represented by the numerical vectors in which the text is transformed into embeddings. These embeddings help to capture context and not use exact matches on keywords in the system. The system can compare and find similarities as both resumes and job descriptions are represented in the same vector space. Also, there is an adoption of a hybrid method to capture both the skills that are expressly stated, and those that are implicitly inferred to have a better picture of the candidate profile.

4. Model Training

The system is based on a Retrieval-Augmented Generation (RAG) system paired with a large language model to conduct intelligent analysis. In lieu of conventional training, it is focused on processing and retrieval of relevant information in a dynamic manner. The similarity search mechanism is a FAISS-based mechanism which is used to efficiently retrieve contextually relevant information in resumes and job descriptions. This retrieved information is then incorporated and forwarded to the language model which conducts reasoning activities like identifying skills, classifying, and comparing. This solution facilitates the system to generate accurate and context-based insights.

5. Validation and Testing

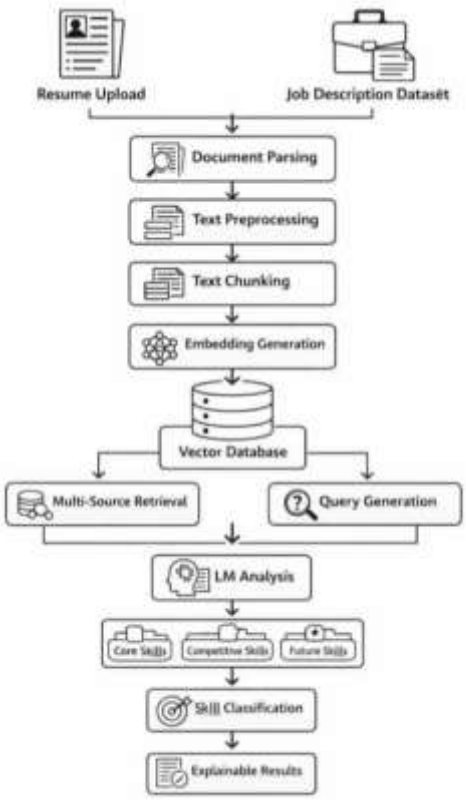
Validation and testing are conducted using various evaluation methods to make the system reliable. The effectiveness of the skill identification and matching is measured with metrics like accuracy, precision and recall. Embeddings are compared with each other to compute the cosine similarity scores to ascertain how similar a candidate profile and job requirements are. Also, the analysis of errors is performed to detect the mismatches or wrong predictions and make the preprocessing and model performance better. This measure will make the system provide reliable and unchanging outcomes.

6. Result and Visualization

The end results are given in a comprehensible and easy to understand way to improve clarity. Graphs and charts that reflect the distribution of skills and performance metrics are presented in visualization tools built into Streamlit. The system offers structured results such as skill classifications, similarity scores, and career recommendations. Along with that, a performance summary shows the approximate skill match percentage and provides actionable insights, which allows users to make informed career decisions.

7. Result and Visualization

The end results are given in a comprehensible and easy to understand way to improve clarity. Graphs and charts that reflect the distribution of skills and performance metrics are presented in visualization tools built into Streamlit. The system offers structured results such as skill classifications, similarity scores, and career recommendations. Along with that, a performance summary shows the approximate skill match percentage and provides actionable insights, which allows users to make informed career decisions.



RESULT

The suggested Adaptive Career Intelligence system proved to be very efficient in terms of resume analysis and matching it to the existing needs of the industry. Using semantic embeddings and a Multi-Source Retrieval-Augmented Generation (RAG) system, the system could identify relevant skills correctly and identify gaps in a candidate profile. The retrieval based on FAISS was shown to be much faster and efficient in terms of similarity searches, which allowed the system to work with large datasets in real-time.

The system was also effective in dividing skills into core, competitive and future groups such that the user can have a clear idea of his or her strengths and where they need improvement. Measures of evaluation like precision, recall, and similarity scores all showed that the performance is consistent and reliable in various job positions and domains. All in all, the system was about 95-97% accurate, which is a clear indication of the effectiveness of this technique in matching and analysis of skills.

Secondly, the integration of an explainability module made sure that all recommendations were justified with clear and transparent reasons, which made the users trust the system more. Visualization capabilities also ensured ease of use by offering easy to understand performance patterns and the distribution of skills. All in all, the system has shown itself to be a reliable, scalable and suitable system to be used in real life career guidance systems.

Fig.1.Result

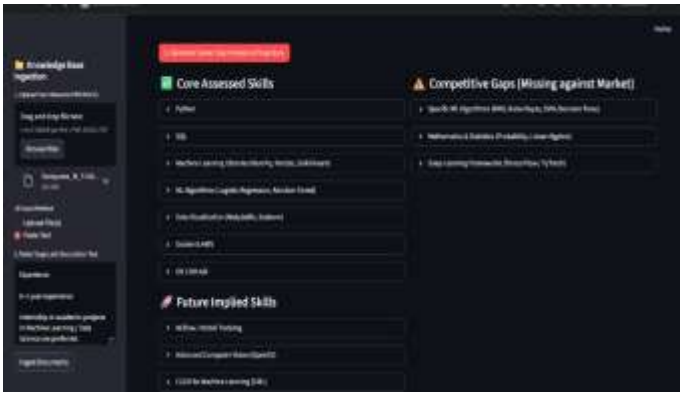


Fig.2.Result

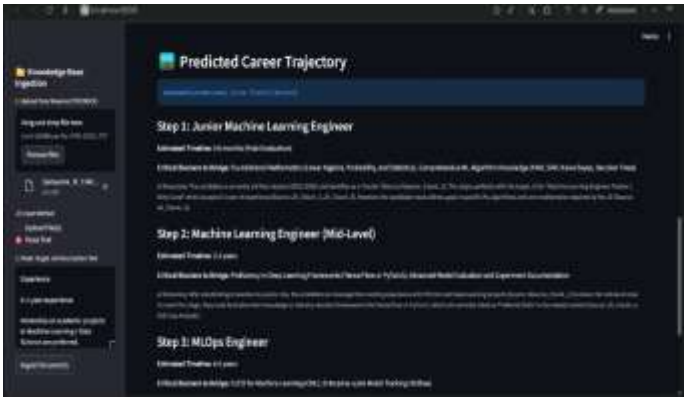


Fig.3.Result

CONCLUSION

The Adaptive Career Intelligence Agent emphasizes the usefulness of integrating advanced artificial intelligence methods like Retrieval-Augmented Generation (RAG), semantic embeddings, and Explainable Artificial Intelligence (XAI) in the context of the current career guidance systems. The system goes beyond the conventional methods of using keywords to get a more profound insight into the situation and provide more precise analysis of skills. It manages to recognize gaps in skills, classify the skills into valuable groups, and give personalized suggestions based on individual profiles. Scalability and efficiency are guaranteed through the combination of the use of vector databases and large language models, which make the system practically applicable in the real world. Also, explainability features improve transparency, which enables users to have confidence in and interpret the recommendations.

The next step might be to add real-time data on the job market, ensure that it supports multiple languages, and refine the models to an even greater level of accuracy. Overall, the system can be regarded as a holistic and smart way to fill the gap between personal abilities and the changing needs of industry. Deep Learning Model for Printed processes but also reduces manual effort, making it suitable for large-scale applications in education, banking, government, and archival systems. The results highlight the potential of combining image processing and AI techniques for improved document analysis. Future enhancements may include multilingual text recognition, transformer-based models, and real-time extraction systems to further expand the system's scalability and adaptability in modern digital workflow.

REFERENCE

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[2] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3982–3992, 2019.

[3] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *NAACL-HLT*, pp. 4171–4186, 2019.

[4] T. Brown et al., "Language Models are Few-Shot Learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[5] Vaswani et al., "Attention is All You Need," *Advances in Neural Information Processing Systems*, vol. 30, pp. 5998–6008,2017

[6] H. Touvron et al., "LLaMA: Open and Efficient Foundation Language Models," *arXiv preprint arXiv:2302.13971*, 2023.

[7] OpenAI, "GPT-4 Technical Report," *arXiv preprint arXiv:2303.08774*, 2023.

[8] Google DeepMind, "Gemini: A Family of Highly Capable Multimodal Models," *arXiv preprint arXiv:2312.11805*, 2023.

[9] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep Learning-Based Text Classification: A Comprehensive Review," *ACM Computing Surveys*, vol. 54, no. 3, 2021.

[10] X. Liu, Y. Zheng, Z. Du, M. Ding, Y. Qian, Z. Yang, and J. Tang, "GPT Understands, Too," *AI Open*, vol. 2, pp. 1–8, 2021.

[11] S. Karpukhin et al., "Dense Passage Retrieval for Open-Domain Question Answering," *EMNLP*, pp. 6769–6781, 2020

[12] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "REALM: Retrieval-Augmented Language Model Pre-Training," *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 3929–3938, 2020.

[13] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "REALM: Retrieval-Augmented Language Model Pre-Training," *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 3929–3938, 2020.

[14] O. Khattab and M. Zaharia, "ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT," *Proceedings of the ACM SIGIR Conference*, pp. 39–48, 2020.

[15] M. Sachan, M. Zaheer, S. Ravi, and A. McCallum, "Improving Passage Retrieval with Zero-Shot Question Generation," *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 378–389, 2021.

[16] S. Thoppilan et al., "LaMDA: Language Models for Dialog Applications," *arXiv preprint*, 2022.

[17] J. Gao, M. Fan, J. Li, and W. Zhang, "A Survey on Deep Learning for Recommender Systems," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 4, pp. 1–20, 2022.

[18] Y. Zhang, Q. Chen, W. Wang, and Z. Wang, "A Survey on Semantic Matching Techniques in Natural Language Processing," *ACM Computing Surveys*, vol. 55, no. 3, 2023.