ViaClaro:
A Career Coach

Supervisor: Dr.Ayesha Enayat
CS, DSSE
Karachi, Pakistan

Batool Ali Akbar
CS, DSSE
Karachi, Pakistan
ba07612@st.habib.edu.pk

Abeeha Zehra
CS, DSSE
Karachi, Pakistan
az07728@st.habib.edu.pk

***Abstract*—ViaClaro is an AI-driven career coaching system designed to address the challenges of career planning and job matching. Leveraging Machine Learning and Natural Language Processing, it predicts career paths and recommends jobs tailored to users' resumes. Despite limitations such as unbalanced datasets, the absence of labeled data, and computational constraints, ViaClaro achieves notable performance, with DistilBERT delivering 99% accuracy in job recommendations and Artificial Neural Networks achieving an f1 score of 0.80 for career path predictions. These results highlight its potential as a scalable and personalized solution, laying the foundation for future improvements in AI-powered career guidance.**

I. Introduction

The intricacies of career planning and job finding are quite daunting, especially for the individuals new to the job market. Individuals often struggle with effectively planning their career trajectory and locating suitable job opportunities. Even though traditional career counselling is in place, it can be quite expensive, time-consuming, and may not provide personalized support. To address these challenges, there is a need for more a accessible, efficient, and customized career support solution.

Many job seekers, particularly those starting out, struggle with aligning their abilities and interests to potential career paths and relevant job opportunities. Our solution, ViaClaro is an AI tool that tries to overcome these limitations by leveraging advanced technologies like Machine Learning, Natural Language Processing, and Large Language Models. This tool will suggest potential career paths and relevant job opportunities based on user's resume and will give a brief on how to approach the said career path.

II. Literature Review

The Resume-Job Matching and Person-Job Fit problem is a persistent issue in the job market, where candidates struggle to find relevant opportunities, and employers face challenges in selecting qualified candidates. AI-driven approaches have been proposed to address this, ranging from resume enhancement and skill extraction using large language models (LLMs) to encoder-decoder models mapping resumes to suitable jobs. Additionally, some research has extended this concept to career path finding, using AI to help individuals identify potential career trajectories based on their skills and experience, thus offering both job matching and long-term career guidance.

One study [1] focused on competence-level prediction and context-aware models, which classified resumes by skill levels and matched them to job descriptions. This approach employed preprocessing methods like section trimming, chunk segmenting, and pruning to remove irrelevant content, ensuring that the transformer-based models could effectively focus on meaningful sections of resumes. Among the models tested by the paper, the chunk segmenting approach, combined with section encoding, achieved the highest accuracy (79.2%) for resume-job description matching, showcasing the effectiveness of processing resumes in manageable segments.

Another research [2] delved into skill extraction using large language models (LLMs), leveraging advanced prompting techniques such as extraction-style, NER-style, and dataset-specific prompts. These methods allowed the LLM to extract explicit and implicit skills from resumes and job descriptions effectively. Preprocessing involved cleaning and integrating diverse datasets across multiple languages to create a robust training set. The results highlighted the importance of precise demonstrations and dataset-specific prompts, improving F1 scores by 20-28% depending on the prompting strategy.

A third study [3] introduced GAN-based frameworks for resume refinement and job recommendations. The approach utilized a generative adversarial network (GAN) to enhance low-quality resumes by aligning them with high-quality examples. The framework incorporated interactive resume completion, where LLMs refined resumes based on user-provided data and job portal interactions. This method significantly improved job recommendation accuracy, especially for users with limited data in their resumes. The GAN's iterative refinement process ensured the generated resumes closely resembled high-quality standards, addressing gaps in traditional resume-building techniques.

A study [4] on interpretable person-job fitting approaches used classification and ranking techniques to evaluate resume suitability for job descriptions. This research introduced seven

feature extraction methods, such as document layout analysis, semantic feature extraction, and skill matching, to quantify the alignment between resumes and job requirements. Preprocessing included parsing resumes into structured components using heuristics and metadata, while features like TF-IDF and doc2vec were employed to capture textual and semantic relationships. Random Forest models achieved the best results, with a precision of 93.7% and an F1 score of 84.3%.

Another notable study [5] explored implicit skill extraction and job recommendation systems using document embedding techniques. This approach combined explicit skills extracted from resumes with implicit skills derived from similar job descriptions. Preprocessing involved training a doc2Vec model on 1.1 million job descriptions to identify and extract skills from the top 10 similar descriptions for each job posting. The bipartite graph matching algorithm ensured effective CV-job mapping, achieving a 20% improvement in recommendation performance when implicit skills were included.

Lastly, the Work Experience Enhanced Person-Job Matching (WEPJM) model proposed by [6] incorporated career-path-aware learning to improve resume-job alignment. By segmenting resumes and job descriptions into statements and work experience components, the model used Bi-LSTM networks on top of a pre-trained BERT encoder to extract semantic features. Auxiliary tasks such as career path identification and reconstruction further refined the model's understanding of candidate preferences, achieving an accuracy of 85% and demonstrating robust performance even for candidates with limited work experience.

III. Methodology

A. Dataset

This project uses two major datasets. The first is the resume dataset which is a combination of Resume and UpdatedResume. Both the resume datasets are csv format files with 2484 and 962 data entries respectively. The first CSV file, titled "Resume" consists of four columns: ID, resume in string format, resume in HTML format, and the career category or domain of the resume. The second CSV file named "UpdateResumeDataset" has two columns: the resume in string format and the career category. We have combined these two datasets to create a dataset of 3446 entries containing two columns: resume represented as a string and its corresponding category or career domain. The categories are HR, Designer, Information Technology, Teacher, Advocate, Business Development, Healthcare, Fitness, Agriculture, BPO, Sales, Consultant, Digital Media, Automobile, Chef, Finance, Apparel, Engineering, Accountant, Construction, Public Relations, Banking, Arts, Aviation, Data Science, Web Designing, Mechanical Engineer, Health and Fitness, Civil Engineer, Java Developer, Business Analyst, SAP Developer, Automation Testing, Electrical Engineering, Operations Manager, Python Developer, DevOps Engineer, Network Security Engineer, PMO, Database, Hadoop, ETL Developer, DotNet Developer, Blockchain, and Testing. The data has been split into an 80 to 20 ratio for training and testing. Figures 1 and 2 show the distribution of these categories in the training and testing dataset respectively.

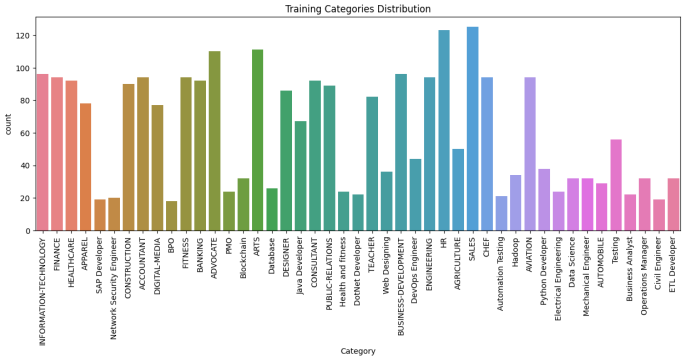

Fig. 1: Training Set Categories

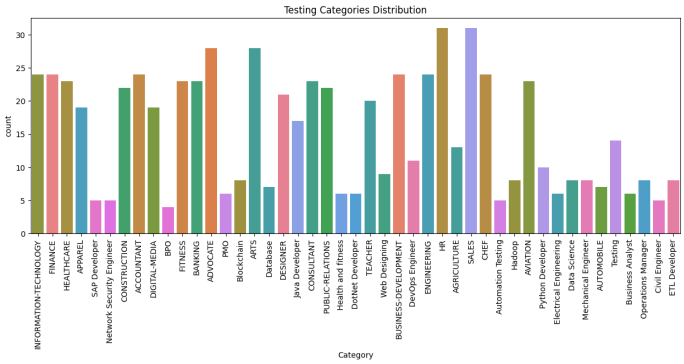

Fig. 2: Test Set Categories

The second data set is a csv file downloaded from Kaggle containing 1.6 million rows and 23 columns. Each row is a different job posting with an ID, Experience, Qualification, Salary, Location, Country, Longitude, Latitude, Work Type, Company Size, Job Posting Date, Preference, Contact Person, Contact, Job Title, Role, Job Portal, Job Description, Benefits, Skills, Responsibilities, Company and Company Profile for every job ID. Because of the limited computation power, we had to limit our dataset to the first 2000 job postings. The distribution of job dataset can been seen in Figure 3.

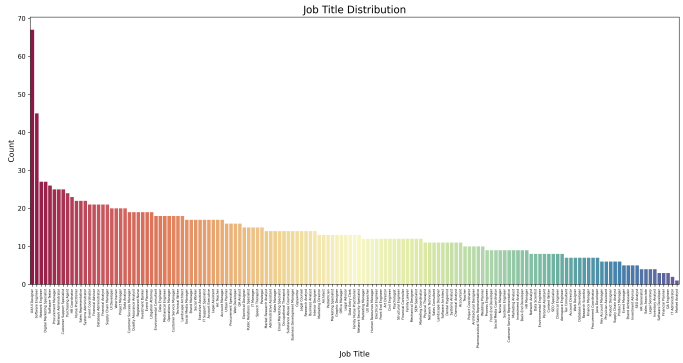

Fig. 3: Job Data Distribution

B. Career Path

The first element to our project is to suggest top 5 career paths to the user which was implemented as the baseline of this project. Our primary approach was to treat this as a classification problem where each category is a separate class. This part of the project uses only the resume data set which was split into a 80 to 20 ratio as mentioned in section III A. We started with the pre-processing of this data. The pre-processing of both the columns is done slightly different. For the first column, we first use the re library to remove stop words, links, special characters, and extra spaces and then create a vocabulary and tf-idf vectors for each word. For the second column, we directly assigned an index to each category.

Once the data was ready, we used it to train 3 models. For this, we used built-in models provided by the sklearn library. Each of these models are trained and saved in their corresponding pkl files.

The first model is the Artificial Neural Network. An Artificial Neural Network (ANN) is a machine learning model that mimics the brain's structure to identify complex patterns. The ANN typically takes encoded word vectors as inputs and passes them through multiple layers of interconnected neurons. For our project, the model is given the tf-idf encoded vectors as an input and it predicts the probabilities of each category. The model is set to have 1 hidden layer of 100 neurons. The ReLU activation function is applied to introduce non-linearity, allowing the network to capture more complex patterns, while the Adam optimizer is used for efficient training.

The second model is a Multi-nomial Naive Bayes Model which is a classification algorithm that applies Bayes' Theorem, assuming that features are conditionally independent given the class. In predicting career pathways from resumes, the algorithm uses the frequency of words to calculate the likelihood of each profession. It uses the training data to predict which words are the most indicative of certain professions and assigns probabilities accordingly.

The third model is the Logistic Regression which is a linear classification algorithm that predicts the probability of a data point belonging to a specific class. In the current context, it works by fitting a linear model to the training data, where each resume is represented by a feature vector of tf-idf values. The algorithm estimates the likelihood that a resume belongs to each profession category based on these features.

C. Job Recommendation

The second element of our project is to recommend the top five jobs to the user based on their resume. To achieve this, we developed a job recommendation system that suggests jobs by comparing the similarity between the user's skills and the skills required for each job. Before delving into the models used, it is essential to first discuss the pre-processing steps undertaken.

As outlined in Section II, breaking the resume into smaller, manageable chunks and segments yielded better results than processing the entire resume as one unit. Inline with this approach, we focused on extracting the qualification (education) and skills sections from the resume. Following this extraction, we performed data cleaning and normalization by removing non-ASCII characters and replacing multiple white spaces with a single space.

Similarly, for the job descriptions, we extracted relevant information such as experience, qualifications, job responsibilities, and skills. After extracting these components, we applied similar cleaning and normalization steps, including the removal of URLs, email addresses, punctuation, and non-alphanumeric characters.

With the data cleaned and normalized, we proceeded to create embeddings for both the resumes and the job descriptions. For generating these embeddings, we utilized two pre-trained models: BERT and DistilBERT, developed by Google and Hugging Face, respectively. BERT (Bidirectional Encoder Representations from Transformers) utilizes bidirectional context to produce deep, contextual representations of text, making it highly effective for a variety of natural language processing (NLP) tasks. DistilBERT, on the other hand, is a smaller and faster variant of BERT that uses knowledge distillation to reduce the model's size while retaining much of its performance.

After importing these models, we used their tokenizers to process each sentence by converting it into tokens. These tokens were then passed through the models to generate embeddings. This process was repeated for both the resumes and the job descriptions. Once the embeddings were generated, we calculated the cosine similarity between each job description and the resume. Based on these similarity scores, the top five job recommendations were selected for

each resume. The working for BERT and DistilBERT can be seen in Figures 4 and 5 respectively.

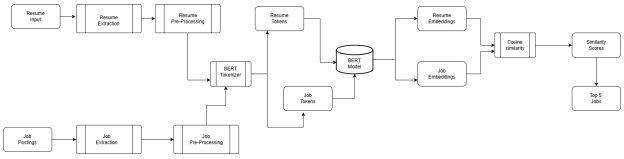

Fig. 4: Job Recommendation Using BERT

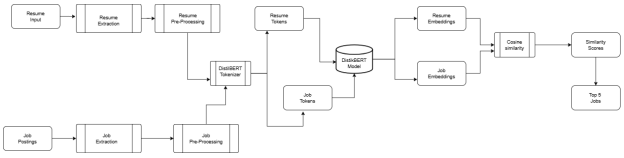

Fig. 5: Job Recommendation Using DistilBERT

D. Front-End Integration of Career Path and Job Recommendation

This section details the implementation of career path and job recommendation features within an interactive interface. The system allows users to upload their resumes as input, which undergoes a comprehensive process of data extraction and pre-processing to ensure accuracy and consistency. For career path recommendations, an Artificial Neural Network (ANN) analyzes the user's qualifications and skills to suggest potential career trajectories tailored to their profile. Once career paths are presented, users can select a specific path of interest. Upon selection, the system utilizes OpenAI's GPT-3.5-turbo model to generate personalized advice and actionable guidance, including the skills to acquire, and steps to take to succeed in the chosen career path.

For job recommendations, the system leverages the pre-trained DistilBERT model to match the user's skills with job descriptions, providing a list of personalized job opportunities.

This dual functionality ensures a seamless experience by integrating advanced AI techniques, enabling users to explore both immediate job options and long-term career development strategies through an intuitive and interactive interface as seen in Figure 6.

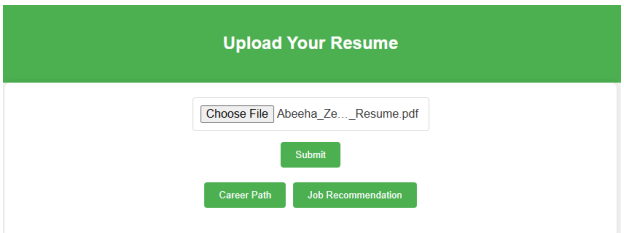

Fig. 6: Front End

IV. Experiments and Results

For the evaluation of all five models, we used a confusion matrix and calculated the recall, precision, accuracy and f1 scores for each model.

The results for the first part of the project i.e. predicting the career path are mentioned in Table 1, 2 and 3:

Evaluation	Score
Accuracy	0.75
Precision	0.82
Recall	0.80
F1 score	0.80

Table 1: Artificial Neural Network

Evaluation	Score
Accuracy	0.57
Precision	0.58
Recall	0.51
F1 score	0.51

Table 2: Multi-Nomial Naive Bayes

Evaluation	Score
Accuracy	0.76
Precision	0.82
Recall	0.80
F1 score	0.79

Table 3: Logistic Regression

For the second part, we did not have the ideal data meaning there was no direct mapping between jobs and resumes to serve as a basis for performance evaluation. Therefore, we devised an alternative method for performance assessment. We used the resume category and the job title of the posted job as a proxy for matching. Similarity between these two elements was calculated, and if the similarity score exceeded a threshold of 0.7, it was considered a match.

While this approach lacked ground truth, we sought to establish a more reliable comparison by identifying overlapping skills between the job and resume. We calculated the intersection of job skills and resume skills, and if an overlap was found, this was treated as a ground truth. This allowed us to calculate the True Positives, True Negatives, False Positives, and False Negatives, and use them to evaluate model performance.

The performance evaluation of both models was conducted using key metrics, including accuracy, precision, recall, and f1 score, as presented in Tables 4 and 5.

Evaluation	Score
Accuracy	0.93
Precision	1.00
Recall	0.93
F1 score	0.96

Table 4: BERT

Evaluation	Score
Accuracy	0.99
Precision	1.00
Recall	0.99
F1 score	0.99

Table 5: DistilBERT

V. Discussion

The graph in Figure 7 depicts a comparison between the three models used for career path prediction.

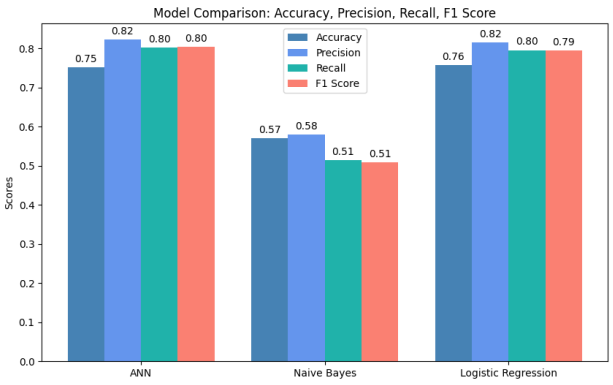

Fig. 7: Comparison between ANN, Multi-Nomial Naive Bayes and Logistic Regression

From Figure 7, it can be seen that Logistic Regression achieved an accuracy of 0.76, with a precision of 0.82, recall of 0.80, and an f1 score of 0.79. While its accuracy is slightly higher than that of the ANN, the model falls behind in the f1 score, which reflects its balance between precision and recall. This slight dip suggests that Logistic Regression might not be as effective in handling the nuances of classifying the most relevant career paths, even though it provides slightly better overall accuracy.

ANN performed slightly better in terms of the f-1 score, with an accuracy of 0.75, precision of 0.82, recall of 0.80, and an f-1 score of 0.80. Although it has marginally lower accuracy, its higher f1 score indicates that ANN strikes a better balance between precision and recall. This improvement suggests that ANN is better at capturing complex relationships in the data, potentially leading to more accurate predictions of the top career paths.

In contrast, Naive Bayes lagged significantly, recording an accuracy of 0.57, precision of 0.58, and recall and an f1 score of 0.51. Its under-performance can be attributed to its assumption of feature independence, which does not adequately reflect the inter-dependencies present in resume data, resulting in a higher number of mis-classifications.

Thus, we can conclude that ANN is the more effective model for this classification task. Despite Logistic Regression having slightly higher accuracy, ANN's superior F1 score indicates a better balance between precision and recall, which is crucial for recommending the top career paths based on resume input. The ability of ANN to capture non-linear relationships in the data makes it more suitable for this task, as it can better account for the complex patterns present in resume features and career categories. Therefore, ANN proves to be the better choice for predicting career pathways from resumes.

The graph in Figure 8 shows a comparison between the two models used for job recommendation. It can be seen that there is a minimal differences between the two models as DistilBERT is essentially a variant of BERT. However, DistilBERT demonstrated superior performance, achieving an accuracy of 0.99, compared to BERT's accuracy of 0.93. Similarly, DistilBERT attained an f1 score of 0.99, while BERT's f1 score was 0.96.

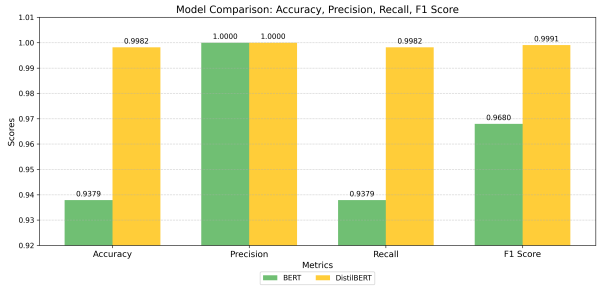

Fig. 8: Comparison between BERT and DistilBERT

A primary observation here is that DistilBERT outperformed BERT, despite both models being based on the same transformer architecture. The primary reason for this improvement lies in DistilBERT's design as a smaller, more efficient version of BERT. By using knowledge distillation, DistilBERT retains much of BERT's performance while significantly reducing its size, leading to faster processing and potentially more accurate results in large-scale tasks.

This efficiency in DistilBERT likely contributed to its superior performance. The smaller model size allows for quicker inference times and less computational overhead, which may help reduce overfitting and improve generalization. As a result, DistilBERT was able to achieve an accuracy of 99%, outperforming BERT's 93%, making it a more suitable choice for the job recommendation task.

VI. Conclusion and Future Work

ViaClaro successfully demonstrates the potential of AI in career path prediction and job recommendation by leveraging models like DistilBERT and Artificial Neural Networks. Despite its effectiveness, the system has limitations that affected its performance and scalability. The imbalance in the dataset, with insufficient resumes and job postings across all categories, introduced challenges in ensuring fair and accurate predictions. Additionally, the lack of labeled data necessitated the creation of custom evaluation metrics, which, while useful, limited the ability to benchmark performance against established standards. Furthermore, the computational expense of generating embeddings constrained the analysis to 2000 job postings, reducing the breadth of job recommendations. Moreover, we integrated GPT-3.5-turbo to get more information on the selected career path. While the information given is correct, it might not be relevant to what the user is looking for since GPT has no context of the resume.

Future work can address these limitations by utilizing properly labeled datasets for more rigorous evaluation. Expanding data sources, such as integrating information from social media or tracking job performance, could enhance recommendation accuracy. Furthermore, incorporating user feedback on preferences, goals, and experiences and incorporating user activity on different job portals will further refine the system's ability to deliver personalized and actionable career guidance. These advancements will ensure ViaClaro evolves into a more robust and comprehensive career coaching tool.

References

[1] C. Li, E. Fisher, R. Thomas, S. Pittard, V. Hertzberg, and J. D. Choi, "Competence-Level Prediction and Resume & Job Description Matching Using Context-Aware Transformer Models," *Proc. 2020 Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 679-688.

[2] K. Nguyen, M. Zhang, S. Montariol, and A. Bosselut, "Rethinking Skill Extraction in the Job Market Domain using Large Language Models," in *Proc. 1st Workshop Natural Lang. Process. Human Resources (NLP4HR 2024)*, St. Julian's, Malta, 2024, pp. 27-42. Association for Computational Linguistics.

[3] Y. Du, D. Luo, R. Yan, X. Wang, H. Liu, H. Zhu, Y. Song, and J. Zhang, "Enhancing Job Recommendation through LLM-Based Generative Adversarial Networks," in *Proc. 38th AAAI Conf. Artif. Intell. (AAAI-24)*, 2024, pp. 8363-8370. Association for the Advancement of Artificial Intelligence.

[4] M. A. Menacer, F. B. Hamda, G. Mighri, S. B. Hamidene, and M. Cariou, "An interpretable person-job fitting approach based on classification and ranking," in Proc. 4th Int. Conf. on Natural Language and Speech Processing (ICNLSP 2021), Trento, Italy, 2021, pp. 130–138.

[5] A. Gugnani and H. Misra, "Implicit skills extraction using document embedding and its use in job recommendation," Proc. AAAI Conf. Artif. Intell., vol. 34, no. 08, pp. 13286–13293, Apr. 2020.

[6] Z. Gong, Y. Song, T. Zhang, J.-R. Wen, D. Zhao, and R. Yan, "Your career path matters in person-job fit," Proc. AAAI Conf. Artif. Intell., vol. 38, pp. 8427–8435, 2024.