AIRM: A New AI Recruiting Model for the Saudi Arabia Labor Market

Abstract. One of Saudi Arabia's vision 2030 goals is to keep the unemployment rate at the lowest level to empower the economy. Research has shown that a rise in unemployment has a negative effect on any countries gross domestic product. Artificial Intelligence is the fastest developing technology these days. It has served in many specialties. Recently, Artificial Intelligence technology has shined in the field of recruiting. Researchers are working to invest its capabilities with many applications that help speed up the recruiting process. However, having an open labor market without a coherent data center makes it hard to monitor, integrate, analyze, and build an evaluation matrix that helps reach the best match of job candidate to job vacancy. A recruiter's job is to assess a candidate's data to build metrics that can make them choose a suitable candidate. Job seekers build themselves metrics to compare job offers to choose the best opportunity for their preferred choice. This paper address how Artificial Intelligence techniques can be effectively exploited to improve the current Saudi labor market. It aims to decrease the gap between recruiters and job seekers. This paper analyzes the current Saudi labor market, it then outlines an approach that proposes: 1) a new data storage technology approach, and 2) a new Artificial Intelligence architecture, with three layers to extract relevant information from data of both recruiters and job seekers by exploiting machine learning, in particular clustering algorithms, to group data points, natural language processing to convert text to numerical representations, and recurrent neural networks to produce matching keywords, and equations to generate a similarity score. We have completed the current Saudi labor market analysis, and a proposal for the Artificial Intelligence and data storage components is articulated in this paper. The proposed approach and technology will empower the Saudi government's immediate and strategic decisions by having a comprehensive insight into the labor market.

Keywords: Recruiting, Job seeker, Artificial Intelligence, Natural Language Processing, Clustering Algorithms, Recurrent Neural Network.

1 Introduction

Recruiting is critical to any organization's success. It is a challenging task with a long process. Nowadays, recruiters must use an effective HR (Human resource) system to select from a pool of applicants the most appropriate candidates that will also focus on emphasizing employee retention [1]. There are several features that recruiters should consider when making a decision. Furthermore, these features will build matrices that can give an indicator value. Features include, for example, the number of vacant positions, the number of applications required, the level of education, knowledge, skills, years of experience, abilities, preferences, values, and more [2].

On the other hand, many job seekers join the labor market – from new graduates to those looking for better or different opportunities. As the market is frequently changing, job seekers are often looking for better positions. The job seekers' activities as a self-regulated mission involve building a CV, finding a career path, finding the best place to apply to, preparing for an interview, comparing job offers, and more. All activities contain features that will build matrices that can give an indicator value. Most research focuses on the recruiters to find the best candidates but not that much attention has been given to the job seekers' side.

It is concluded from Okun's Law that the rise in unemployment in the US affects the gross domestic product (GDP). According to Okun's Law, a one percent increase in unemployment causes a 2% fall in GDP [3]. According to the General Authority for Statistics, the Saudis' unemployment rate in Saudi Arabia for the second quarter of 2019 is (12.3 %). And this is a high percentage for a rich and developing country with a population that does not exceed more than 20 million. The importance of helping to reduce unemployment is clear here, and the Saudi Vision 2030 addresses this issue.

Saudi Vision 2030 is a plan that was announced on April 25, 2016, and coincides with the date set for announcing the completion of the handover of 80 government projects. The plan was organized by the Council of Economic and Development Affairs. And jointly achieved by the public, private and non-profit sectors [4].

The Kingdom of Saudi Arabia is making tremendous efforts in this field. However, there are many obstacles in the Saudi labor market, each with different dimensions. It is not within the limits of this paper to discuss them all. However, it might be helpful to review some of them, to get an overview of these Saudi market obstacles. For example, market regulations, the wage gap between national and expatriate workers, most nationals employed are in the public sector, not enough well-paying and productive jobs for the young and growing population, just some of the obstacles [5]. To overcome such obstacles to efficient recruitment, a national data center to store and integrate data that integrates Artificial Intelligence (AI) models to accurately analyze and predict trends could be very useful.

To test this hypothesis, this paper proposes an AI model as a matching engine that serves both recruiters and job seekers, using labor market big data. The model will utilize text mining models, natural language processing, and clustering algorithms to extract and analyze relevant information that is integrated, from different sources, into the national data lake (DL). This approach will allow mapping between job supply and demand in the labor market, which is too complex to achieve manually.

This paper is organized as follows. Section 1 provides a brief description of Saudi Arabia's labor market. Section 2 highlights the DL (i.e. the data repository) and the AI model's techniques that will be used in the proposed solution: Clustering Algorithm, Natural Language Processing, and Returned Neural Network. Section 3 describes the solution proposed. Finally, Section 4 presents the conclusions of this study.

1.1 Background Research

A labor market can be defined as a mapping process that forms a mechanism to match demand with supply for employers and employees [6]. The term 'labor market' is criticized by the US Secretary of Labor. He thinks that this term regards employees as being bought and sold, such as any other product, whereas they are unique in several ways. The term labor market refers to supply and demands in this field. Employees sign clear contracts with some tasks and responsibilities. The contract involves employees' rights, salary, allowances, and other detailed constraints binding on both parties [7].

Historical transformations in the labor market has generally been due to changes in labor conditions, and some factors have dramatically changed the characteristics and nature of the Saudi labor market in the past few decades, and both advanced and developing countries. Several jobs have disappeared while new jobs have become available; some are genuinely novel jobs that did not exist until a few years ago. Also, the quantity and quality of skills and qualifications in demand associated because of this new labor market has changed dramatically. All of this might influence the nature of the new labor market.

This research, outlined in this paper, first analyzes the Saudi labor market in-depth, trying to understand the Saudi labor market, national projects that currently support the labor market, the labor market growth, the current recruiting system, and the new establishment of the Saudi data and AI authority. This research poses some interesting questions. With all projects that the government has implemented, why did the situation not improve significantly? How can AI be effectively exploited to improve the current Saudi labor market? What is the best repository to store national data? What are the machine learning (i.e. clustering algorithms) that can be used to build the model? Finally, how to structure the new approach? To try and answer these questions we discuss in more depth the: Saudi labor market, Saudi national projects for the labor market, Saudi labor market growth, the current recruiting system in Saudi, and the Saudi data and AI authority, next. This leads us to propose a new data repository architecture and AI model for matching Saudi job seekers to the Saudi labor market see Section 2 Data Lake and Algorithms.

Saudi Labor Market. The Kingdom of Saudi Arabia is currently witnessing an unprecedented economic transformation, which has affected all government activities and may have the effect of creating new jobs and bringing more Saudi women into the labor market, bearing in mind that one of the economic objectives of the Saudi Vision 2030 is to reduce foreign remittances [8]. The rapid revolution in the labor market might take an unexpected road, where changes in the skills needed for current jobs coincide with the emergence of new jobs, requiring an effective monitoring skills method, which has not been available until now [9].

The Saudi labor market relies heavily on foreign workers, especially in the private sector, for two reasons. The first reason is the massive demand for workers in the oil sector. The second reason is the size of the Kingdom of Saudi Arabia, which needs large infrastructure projects that require temporary workers who work only for the duration of the project and therefore do not provide secure employment opportunities for Saudis. Saudi Arabia has two distinct labor markets with different characteristics with many such workers, one for the Saudis and the other foreign workers [10].

In general, Saudi Arabia's labor market is divided into a government sector that follows the General Organization for the Retirement System and a private sector whose pension system is subject to the social security system. The International Labor Organization (ILO) has defined underemployed as individuals searching for work and available to work more hours but worked fewer hours than their capacity and/or willingness to work [11]. Employment for Saudis is at the forefront of the discussion of the Kingdom's economic policies. However, there is still much research needed to identify solutions that ensure adequate and sustainable jobs for Saudi citizens. Labor market developers aim to find jobs in the private sector for citizens, where the significant difference in both labor rights and the cost of employment between nationals and foreign workers means that employers always favor foreign workers. Stephen points out that the employer's perspective is often missing in the discussion of Saudization[1], and it should be seriously analyzed to identify policies that work on the ground, rather than employer's evasion through 'delusional work' and other fraud techniques on labor market regulations [12].

Abdul Hamid Al Omari is an economic specialist for a Saudi financial agency and one of the well-known economic writers in Saudi Arabia; he reviewed an analytical study of the Saudi labor market and explained no official documented information from the government approval of data sources. The conflicting information about Saudi Arabia's labor market is only due to the multiplicity of semi-official and natural bodies dealing with the employment situation. Additionally, he summarized the most important characteristics of the labor market in Saudi Arabia, which were developed by the Labor Force Council, as follows:

- The lack of adequate data on the Saudi labor market, employment, and unemployment.
- The largest age group in the Saudi population is children and adolescents. This is reflected in the increase in the working-age population.
- Saudi women's contribution to the labor market is low, almost 6.0%. Although a large proportion of Saudi women applicants have university qualifications, there are limited opportunities available to Saudi women.
- The lack of relevance of the current education system to modern developments in Saudi society and the imbalance in the structure and curricula have been revealed by monitoring its scientific level.

From the above point, the first and most important reason for unemployment among the population is the superficial education level. In addition, there is a lack of training or low levels of it, both before and during employment, which may cause staff to lose their jobs, thus contributing to the rise in unemployment.

There are massive differences between foreign workers in terms of qualifications and skills and their professions, as a large proportion of foreign labor is unskilled, and workers hold low-paid occupations and do not require any scientific or technical

[1] Saudization is the newest policy of the Kingdom of Saudi Arabia implemented by its Ministry of Labor and Social Development, whereby Saudi companies and enterprises are required to fill up their workforce with Saudi nationals up to certain levels.

skills. To illustrate, 79.6% of foreign workers fall into this category; only 20.4% are taking up professional [13].

Saudi National Projects for the Labor Market. The problem of unemployment and employment with inappropriate qualifications is not new. Solutions have always been temporary; although the government has implemented many programs to solve this problem, the situation has not improved significantly. National programs are shown below in Table1.

These programs include the HADAF program, one of the mechanisms that contribute to the provision of qualified Saudi cadres, whereby trained, educated young people of both sexes achieve strategic goals, which has social and economic benefits and provides security. Capacity services are focused on the functional linkage between job seekers and private sector employees. The needs of the labor market can be addressed through several channels, such as the site for e-recruitment and re-training and employment centers; as well as applying a system to protect wages, achieve Saudization, the incentive program, the Human Resources program, and the National Labor Observatory Portal [8].

The National Labor Observatory portal is part of the initiatives to stimulate the private sector to expand Saudization. It is also one of the critical national initiatives contributing to improving and developing the market and supporting decision-makers as part of The National Transformation Program, one of the Saudi Vision 2030 initiatives developed to serve its people [14].

The most prominent current products of the National Labor Observatory are indicators of the Saudi labor market, its definition, the formulation, and participants' characteristics. Characteristics include the private sector's social insurance, mobility and job stability, graduates' employment, and establishments that recruit subjects to NETAKAT.

NETAKAT and TAQAT are two initiatives of the Saudi Ministry of Labor. NETAKAT evaluates establishments operating in the Saudi market. The constructs' ranges are classified into four levels: platinum, green, yellow, and red. The originality ranges are divided into two main categories: enterprises with less than ten employed and enterprises with more than ten employed. First, only one citizen must give an enterprise the advantages of a domain initiative, and the second requires different settlement rates.

On the other hand, the TAQAT program offers a range of specialized services provided by the Human Resources Development Fund HRDF to support job seekers by delivering research a database of job seekers from the citizens to choose suitable candidates [14]. The Saudi Ministry of Labor also provides many other programs and initiatives that ensure smooth structural transformation in the Saudi labor market's composition, including developing market control mechanisms, combating concealment and deportation of offenders, developing remittance systems, and protecting wages [10]. There are pressing government movements to overcome labor market problems by launching all these programs and integrating them.

However, full data integration of all of these programs is an obvious problem, that could be solved by developing a national DL to serve as a national repository. **Table**

1. Saudi national projects for the labor market illustrates some of the overlapping characteristic the existing programs mention and highlights the need for a national repository.

Table 1. Saudi national projects for the labor market

The Program	Training	Funding	Job search	Evaluate the workplace
HADAF	√	√		
TAQAT			√	
NETAGAT				√
HAFEZ	√	√		

Saudi Labor Market Growth. In this research, the statistics are obtained from the General Authority for Statistics (GSTAT). The statistics show a significant gap between Saudis and non-Saudis. Out of 9,093,773 employees in the General Organization for Social Insurance in Saudi Arabian (GOSI), 7,157,265 are not Saudis, and only 21.29% are Saudis. These figures are not for employees with low qualifications or insufficient education levels, as they relate to the professions shown in **Table 2**. Out of 923,504 Saudi job seekers, there are 549,851 that hold Bachelor's degrees, as shown in **Table 3**.

Table 2. Education levels

Education levels
Lawyers
Directors and business managers
Specialist professionals
Technical and humanitarian staff
Professional technicians
Clerical occupations
Retail staff
Service occupations
Skilled agricultural occupations
Animal husbandry & fishing
Industrial occupations
Chemical operations
Occupations that support basic engineering.

The total of 2,371,390 employees who are non-Saudis and hold domestic household occupations are not considered in this paper due to the low payment and poor education level [15].

Table 3. Education status and Saudi job seekers

Education Status	Saudi Job Seekers
Illiterate	20,592

Can Read & Write	7,637
Primary	42,860
Intermediate	44,719
Secondary or Equivalent	240,981
Diploma	9,423
Bachelor's degree	549,851
Higher / Diploma	5,082
Doctorate	228
Not Specified	2,131
Total	923,504

From our analysis of the Saudi labor market, we conclude that there are significant problems with an unbalanced labor market where non-Saudis are dominating this labor market. In this paper, we will not try to find the reasons or historical accumulations of the problem as our focus will be to build a model to match candidates with jobs and produce a more Saudi-oriented labor market trend, thus supporting the Saudi government's Saudization policies.

The Current Recruiting System in Saudi. Recruiting data comes from the General Authority for Statistics (GSTAT), who collect their data from the General Organization for Social Insurance (GOSI), the Ministry of Labor and Social Development MLSD, the Human Resources Development Fund HRDF, TAQAT, the Ministry of Civil Service MCS, the National Information Center NIC, the Ministry of Education, King Saud University, and public organizations for technical and vocational training. The data does not include employees in the military sectors and those not registered in the records of GOSI, MCS. Further, the data of the GOSI, MCS, is preliminary. What is worth mentioning is that there is no fully comprehensive link between these entities currently as the data is collected from officially presented entities in paper format.

There is a mismatch in official data on employment in the private sector, which comes from two primary sources: The Department of Statistics and Information and the Ministry of Labor. The inconsistency can be observed by measuring the ratio Saudization in the private sector. The percentage of Saudization in 2014, using the data from the Department of Statistics, was 22.1 percent, but was not more than 15.5 percent for the same period, according to the Ministry of Labor. For ongoing correction of the employment situation in the Kingdom, a statement shared by the Department of Statistics and the Ministry of Labor was issued in February to confirm that Saudi General statistics are the primary source of employment statistics [16]. This mismatch is one of the main reasons for the lack of economic indicators of employment and unemployment in the market, which helps to monitor the performance of the labor market and informs economic policies that aim to correct the current situation.

Due to the Covid-19 pandemic, the unemployment rate among Saudis has risen to 15.4% in the second quarter of 2020. This represents an increase of 3.1 percentage points over the same period in the previous year. From the previous year, the overall unemployment rate (for Saudis and non-Saudis) increased to 9.0%, up 3.4 percentage points from the second quarter. The year 2019 AD and the results of the labor force survey were significantly affected by the effects of the covid-19 pandemic on the Saudi economy. The total labor force participation rate (for Saudis and non-Saudis) is 59.4% during the second quarter of 2020 [17].

Currently, the methods of job search are as follows: prospective employees apply directly to an employer, fill in and send an employment application form by post or electronically; or they ask friends and relatives about job opportunities. Answering published advertisements for official jobs necessitates registration with the Ministry of civil service. People can also register with private employment offices or start a private business. To do the latter, they should apply for a permit or license to start their own business [15]. The public sector depends on promotions in the public sector, depending on its age and length of service [16]. This suggests that the Saudi labor market is an open market to the extent that there are highly diverse ways to apply for jobs in both government and private sectors.

There are some effects of the initiative program of HADAF, including the ability of companies to manipulate the system, which may explain the lack of full commitment by the employees of the institution in two ways:

1. Employing Saudis temporarily is one-way companies circumvent the system and avoid recruitment restrictions on expatriate workers. They improve their image by hiring many Saudis when they need to hire foreign employees or update their work visas. The program attempts to prevent this circumvention by requesting the workers' records at an average of 12 working weeks for Saudi employees. However, there are still reports of companies employing many Saudis with low salaries over a short period.
2. Reducing size to avoid the required Saudization ratios. One way to avoid the penalty is to reduce the number of workers in the company to less than ten to be listed outside the program ranges [18].

Al Omari discusses how companies in Saudi Arabia need to find ways to continue to attract experienced and expatriate Saudi talent to maintain their success. He also suggests companies need to design a work environment desired by employees and offer a more extended period. He further reflects that the Saudi government needs to develop a broader conception of market characteristics and focus on them when analyzing and studying the labor market in Saudi Arabia. Finally, his work is focusing attention on analyzing and delving deeply into the various aspects of these critical characteristics of the Saudi labor market that will help those interested in the market overcome the obstacles of Saudization [13]. Labor market analysts see that labor market problems in Saudi Arabia are concentrated on the dependence on foreigners in the private sector, where the unemployment rate was 12% in 2017 [8].

The imbalance of the labor market in the private sector, which is characterized by its heavy reliance on low-paid expatriate workers (about 80% of the workforce is low-skilled workers with primary or lower education), has exacerbated unemployment among Saudi youth and has contributed to reducing the efficiency and productivity of the private sector. This kind of employment cannot significantly contribute to the transfer of knowledge and the competitiveness of the economy as it depends on its productivity on physical exertion [10]. Unemployment is one of the economic indicators of labor market performance, and it affects families by the loss of their purchasing power, and the nation generally loses their contribution to the economy. Unemployment is also a driver of migration patterns. The problem of poverty and unemployment have always been critical obstacles to economic development [19]. The main conclusion drawn about the impact of the labor market problem on the economy is that the high unemployment rate in countries that are weak in economic growth is not surprising, but it is not expected to occur in a rich country with profitable economic growth like the Kingdom. Therefore, solutions urgently need to be found; as such this paper proposes a model to solve some of these problems.

Restructuring the Saudi economy is a long-term strategic development objective, but it cannot be done in isolation by reforming the labor market only, especially in the private sector, which relies heavily on low-wage and low-skilled expatriate labor. This excessive dependence on this type of employment reduces opportunities for the development of the Saudi economy structure. It also reduces the provision of job opportunities for Saudi citizens. It is understandable that unemployment rates are high in countries that are weak in economic growth, but it is not expected to occur in a rich country with profitable economic growth like the Kingdom.

The Saudi economy is heavily dependent on the oil sector, which does not provide sufficient employment opportunities, neither do other sectors directly related to the petrochemical industry. Also, as the government sector cannot absorb this large number of job seekers, the private sector is the only sector whose growth level can be reflected in the creation of more jobs. Private sector growth in the last ten years has been around 9.7%. Naturally, this growth has been accompanied by an increase in the number of jobs. Indeed, the number of job opportunities in this sector increased by approximately 83%, but Saudi citizens filled only 17% of the private sector posts, and expatriate workers acquired the rest.

The radical solution is to obtain accurate data that can be integrated from all sources and analyzed. In this respect, the Kingdom of Saudi Arabia government has made tremendous efforts towards digitalization, as it has established an authority called the Saudi Data and Artificial Intelligence Authority (SDAIA).

Saudi Data and Artificial Intelligence Authority. Saudi data and AI authority (SDAIA) it is a new establishment is Saudia Arabis started in 2019. It supports the achievement of the Kingdom's Vision 2030 and unleash the Kingdom's capabilities and intending to build a data-based economy. SDAIA works to regulate the data sector and enable innovation and creativity through three arms: The National Data Management Office and the National Information Center, and the National Center for Artificial Intelligence. Unlocking the latent value of data as a national wealth to achieve the aspirations of Vision 2030 by defining the strategic direction for data and AI and supervising its achievement through data governance, providing data-related and forward-looking capabilities, and enhancing them with continuous innovation in the field of AI [20].

The National Data Management Office in SDAIA is building a National Data Bank, which will regulate the injection of stream data flowing from all government agents. The aim is to control the power of data that opens many opportunities and gives a clear national agenda to solve many problems. Using data will pave the way for innovation and achievements, and by managing it well, it will become a valuable source of wealth not only for the Kingdom but also for the world [20].

DLs are emerging as an increasingly popular solution for Big Data at the enterprise level [24]. It has significant advantages over traditional data warehouses. Data scientists, data analysts, and data engineers can access data much easily and faster than would be possible in a traditional data warehouse. Increase the agility and provide more opportunities for them to explore and proof of concept activities.

2 Data Lake and AI Model

Getting all the data in one place will support integrating it in a way that allows data engineering to clean it, then for data scientists to analyze it, and to apply ML algorithms on the data. This section first provides a background of DL and utilization. Second, it highlights the fundamental understanding of ML algorithms needed to build this AI model, such as clustering and NLP.

2.1 Data Lake

James Dixon, first mentioned the concept of a Data Lake (DL) as a data repository in 2010. He stated that a DL manages raw data as it is ingested from multiple data sources. It does not require cleansed data or structured data [21]. A DL is a daring new approach that harnesses the power of big data technology. It is "A methodology enabled by a massive data repository based on low-cost technologies that improve the capture, refinement, archival, and exploration of raw data within an enterprise" [22]. Data are stored in the DL in their original format, whether it is structured, unstructured, or multi-structured. Once data are placed in the lake, it is available for analysis [23]. DLs are often described in the literature with the characteristics illustrated in **Table 4**.

Table 4: Key characteristics of a DL from a technical and business perspective

DL Characteristics	
Business Requirements	Technology Requirements [24]
• DLs are essential for companies and businesses. It gives them a competitive advantage in the data storage domain. The distinct characteristic is that it attracts more attention from business fields instead of academic research fields [23]. • A capability where a business can get raw data, i.e. unchanged data, from different source systems in an enterprise, readily available for analysis [25], [26].	• DLs are a collection of technologies that serve the data's need as a central repository • DLs serve as a cost-effective place to conduct a preliminary analysis of data [27]. • DLs are created to handle large and fast arriving volumes of unstructured or semi-structured data for further dynamic analytical insights. • DL data can be accessible once it is created in the DL.

• A relatively new concept whose definitions, characteristics, and usage is currently more prevalent in web articles than academic papers. • DLs are built on the concept of early ingestion and late processing, and it should be integrated with the rest of the enterprise's IT infrastructure.	• DLs require maintaining the order of the data arrival • DLs will being flexible and task-oriented, data structuring should be implemented only where the DL outflow is the analyzed data for what is necessary. • DLs should handle SQL, NoSQL, OLAP, and OLTP • DLs have a flat architecture, where each data element has a unique identifier and a set of extended metadata tags • A DL forms a vital component of an extended analytical ecosystem. • There should be different possibilities to split data in the DL, i.e. DLs can be partitioned by their lifetime or the type of data. • For DLs partitioned by their type: Raw data, augmented daily data sets, and third-party data. • For DLs partitioned by lifetime: Data that are less than six months old, older but still active data, and archived data [24].

To ensure the effectiveness of a DL architecture, you must keep the following in mind while building and storing data. See **Table 5:** DL architecture build characteristics and data types

Table 5: DL architecture build characteristics and data types

Architecture build Characteristics	**Data Types** [24].
• Capability to expand very large • Flexible policies and governance should be developed according to the need for identification, retention, and data disposition. • DL governance should include an application framework for ▪ Contextualizing data ▪ Advanced metadata management, ▪ Centralized indexing, ▪ Consider the relation between data stored, ▪ Keeping track of data usage ▪ Fully shareable and accessible data, ▪ Shared access is simultaneous, ▪ Access from any device to support the mobile workforce. ▪ Agile analytics.	• Transaction logs • Sensor data • Social media • Document collections • Geo-location • Images • Video • Audio

A typical DL Architecture usually consists of three layers; a data source layer, a processing and storage layer, and a visualization (target) layer, as shown in **Fig. 1** [28].

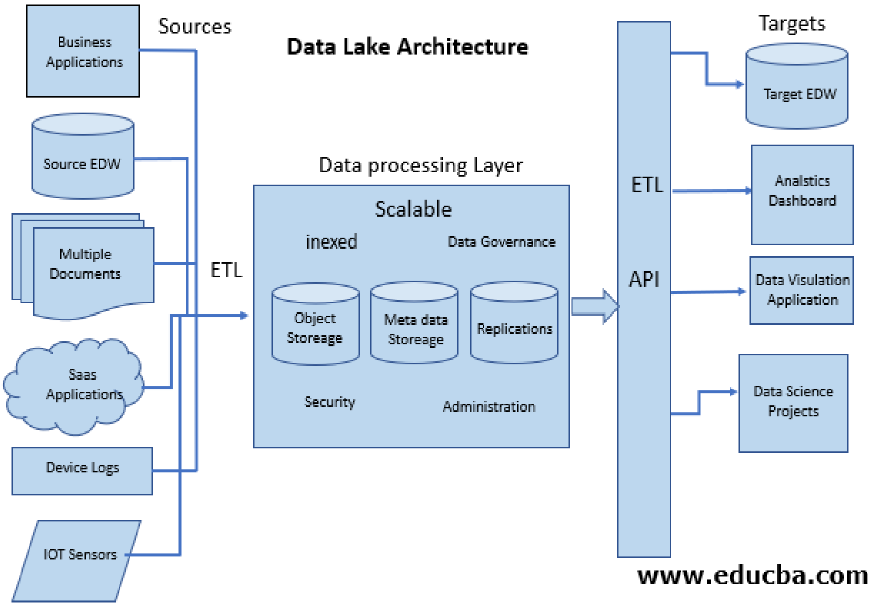

Fig. 1. High-level DL Architecture [29]

There are two well-known types of DLs: a logical DL and a public DL. The source layer can consist of homogenous sources, similar data types or structures, easy to join, and consolidate data, and/or heterogeneous sources, which means different data formats and structures. A method of extracting, transforming, and loading (ETL) is needed to aggregate the raw data from the sources.

The data processing layer, efficiently designed to datastore, metadata store, and the replication for the high availability and to support the security, scalability, and resilience of the data and proper business rules and configurations are maintained through the administration. The DL's target layer (visualization) receives data from the processing layer through an API layer or connectors [30].

Several reference DL architectures are now being proposed to support the design of big data systems. Here is represented "one of the possible" architectures (Microsoft technology-based), as shown in **Fig. 2** [31].

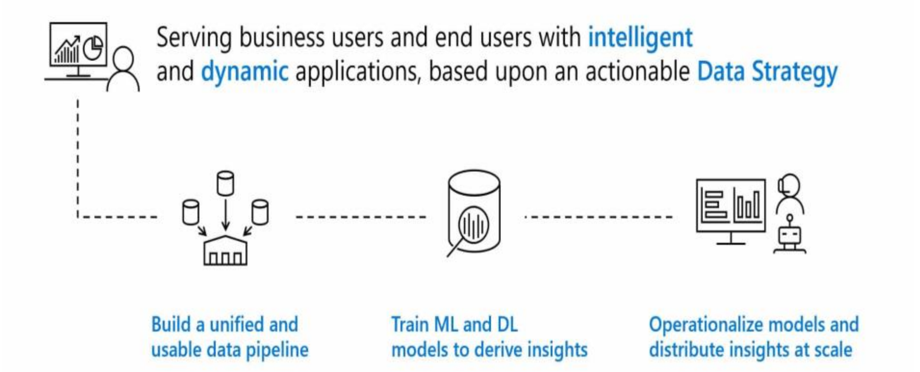
Serving business users and end users with intelligent and dynamic applications, based upon an actionable Data Strategy
Build a unified and usable data pipeline
Train ML and DL models to derive insights
Operationalize models and distribute insights at scale

Fig. 2. High-level Microsoft technology-based [31].

2.2 Algorithms

Data collection and preparation are fundamental for running ML models. Having data in a DL we can store data as it is, not needing to first structure the data, or run different types of analytics at the time of analysis 'schema on read'. Query results are obtained faster using low-cost storage from more sources in less time, and here ML helps to automatically find complex patterns in data. Because ML is able to act without being explicitly programmed this helps decision-makers to take more accurate decisions.

Clustering Algorithms. We can see from **Fig.**1, that we need to process raw data in the DL through the ETL/API process to present useful information to a visualization target whereby a user (job seeker or recruitment agent) will use the result, e.g. use the processed data to match job seeker to job. We need to define suitable processing algorithms that can match and classify data efficiently to do this task. A suitable class of algorithms are Clustering Algorithms [31], which are classified as unsupervised machine learning algorithms. These algorithms desire good at discovering data patterns to build natural groups for similar data points, and are particularly useful if there is no explicit data class to be predicted. Different types of clustering algorithms and choosing a cluster algorithm type depend on the case and the data used, and there is not one single best type for all cases [32]. With our architecture, we will consider the BIRCH clustering algorithm for processing and classifying data.

The BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) is an agglomerative hierarchical clustering algorithm. It is used best with extensive databases. It works by partitioning objects hierarchically using a tree structure and then applies other clustering algorithms to refine the clusters. It iterates by using new input data points and tries to produce the best quality cluster. It finds the best clustering with a single scan of the data. Then it improves the results with a few additional scans [33].

Natural Language Processing − Word Embedding. We have data of type text in our data that we need to process. ML works only with numeric representation. Some of it is free text, e.g. job description, others can be selected from multiple-choice entry, e.g. job level. ML does not accept text input. Here where we need a type of AI model to process text and convert it to numeric representation, for this task we need to exploit Natural Language Processing (NLP).

NLP is a part of the Artificial Intelligence, Computer Science and Linguistic fields of study whose focus is on making computers understand the sentences or words written in human languages [34]. NLP has gained much attraction recently because of the many applications and fields in which it can serve. Word embedding is a subfield of NLP that is concerned with the process of converting words to real numbers to feed them to algorithms, where algorithms as known do not accept text representation. In other words, it is a numerical representation of text which captures the data [35]. The data contains job descriptions and candidate skills that need to be analyzed to enable us to get to the best match. Using prediction-based methods will enable us to analyze and find words of similarity between job specification and job seeker that we need for matching.

NLP prediction techniques are used to generate word embedding, which capture meanings, as well as the semantic relationships of words. Given words from their context, trying to predict new words, using the surrounding words to predict the word of interest. Specifically, NLP prediction is a ML algorithm that has been trained on a large corpus of text data. The training process involves either using a word and trying to predict words that occur in its context or using words in some context to try and predict a specific word of interest. This word embedding model is low in dimensionality. Essentially what this ML algorithm tries to do is to encode each word as a vector of other words. Therefore, keywords from job description will be highlighted and mapped with keywords from candidate skills. There are famous ML algorithms that have pre-generated word embeddings that can be used in models. Examples are Word2Vec, GloVe from Gensim library [35].

Currently, a job recruiter looks for keywords in the job description and then tries to find the similar words in the candidate skills and qualification. The key to our AI model success will be trying to make the model replicate this behavior automatically, and learn from its mistakes. To do that, we will use techniques to find a keyword in job descriptions and in candidate skills and qualifications to map them.

The first technique is TF-IDF where TF stands for term frequency and IDF stands for inverse document frequency. To capture how often a word occurs in a document, as well as how often that word occurs across the entire corpus, both the frequency and relevance of words representing the significance of each word [36]. Capturing how often a word occurs in a document will present the keywords in job descriptions and candidate skills that is needed for mapping.

The second technique is Recurrent Neural Networks (RNN). deep learning can eventually give machines an ability to think, analogous to common sense, which can augment human intelligence. RNN will be used to support the generating of keywords as it proved its power in this field [37].

Search Models Background. There has been some previous research work in the recruiting field, which were mainly focused on linear models. This work tried to find the best candidate for the recruiter but did not consider the complex relationships between features. All previous solutions handled recruiting problems from different dimensions. For example, Ha-Thuc and his team introduced a collaborative filtering approach based on matrix factorization to generate candidates' skill expertise scores, then they utilized them as features in supervised learning and sorted them for normalized discounted cumulative gain NDCG [38], [39].

Another example in this domain is done by Sahin and his group. They proposed a system that is divided into an online system for serving the most relevant candidate results and an offline workflow for updating different machine-learned models [40]. This paper is distinguished from others, where it deals with data resident in The Saudi national data bank where the history of the job seeker is available and can be integrated with the job seeker job profile. Moreover, it works with three layers using a combination of unsupervised learning algorithms. Each layer makes it easier for the next layer from the computational perspective. Finally, it takes the users' preferences and builds weight that is used to prioritize the results.

3 Proposed Solution

This section focuses on our proposed solution for a robust AI model using Python and associated libraries (see section 3.2) to create a proof of concept. Our solution extracts relevant information from available data on both sides, the recruiters and job seekers. Our AI model consists of three layers: the Initial Screening layer, Mapping layer, and a Preferences layer. The three layers work in sequence to match the job seeker with the best job ID, as shown in **Fig. 3**. Our proposed model makes the following requirements:

- A national recruiting platform already exists.
- Data are clean and prepared for analysis.
- All recruiters' and job seekers' data are in the DL.
- Job seekers can be tagged to one job or more.
- Job ID can be tagged to one candidate or more.
- A comprehensive directory of jobs and majors that suit them should be stored in the DL, updated whenever needed.
- Job seekers' data should include personal histories, skills, capabilities, a responsibility that job seekers are willing to take, and accomplishments.

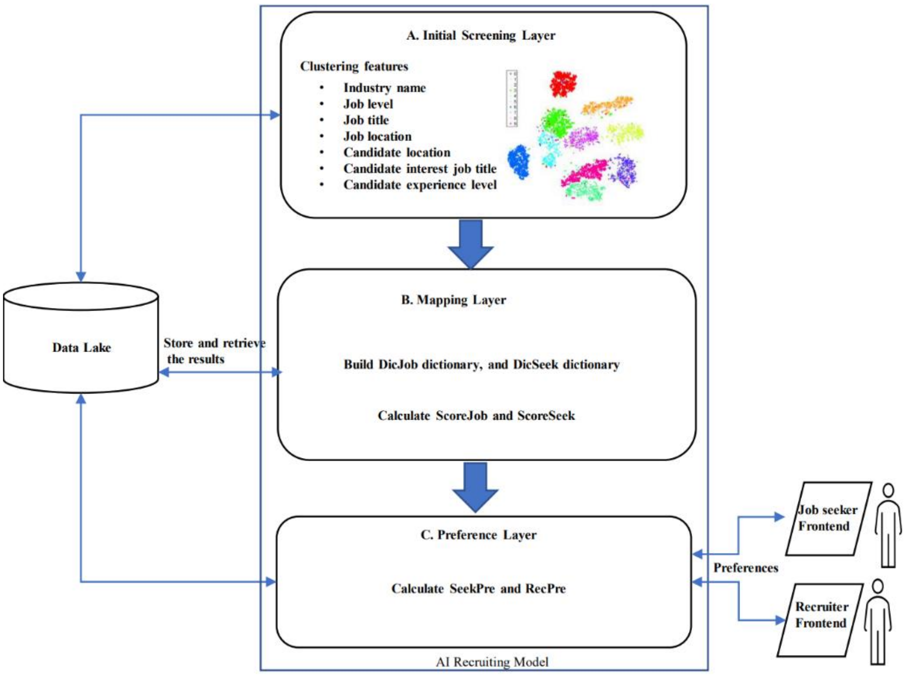
A. Initial Screening Layer
Clustering features
Industry name
Job level
Job title
Job location
Candidate location
Candidate interest job title
Candidate experience level
Data Lake
Store and retrieve the results
B. Mapping Layer
Build DicJob dictionary, and DicSeek dictionary
Calculate ScoreJob and ScoreSeek
C. Preference Layer
Calculate SeekPre and RecPre
Job seeker Frontend
Preferences
Recruiter Frontend
AI Recruiting Model

Fig. 3. AI Recruiting Model

3.1 Initial Screening Layer

This layer will work as a preparation phase. It will use BIRCH to build cluster groups job specializations, which will enable the second layer to treat each cluster specialty separately. This layer's input is from both sides, the recruiter's data and job seekers' data. From the recruiter's data, the industry name, job level, job title, job location, the employment period is full-time or part-time, and gender from the job seekers' side is considered. Other required data includes industry, location, employment period it full-time or part-time, and gender. The AI model will set these groups and their ID. The result will be stored as a data frame in the DL. It will be an iterative stream process that considers the immediate changes from the user profile. This layer will reduce the AI model's computational requirements for the next layer to enable the Mapping layer to work only with the needed group ID.

3.2 Mapping layer

This layer will deal with the job groups that have been clustered and given a cluster ID each for both sides. It will parse the critical skill and responsibilities from the job description as well as parsing the qualification, experience, and responsibilities that the job seeker is willing to take from the job seeker data by using the Python Natral Language Toolkit (NLTK) library for processing the text data from both sides. A dictionary of keywords will be created with a unique ID by using RNN. This will be stored in the DL and retrieved when needed. After that, Word2Vec will be applied to convert the words in the job description and the job seekers qualification to numeric values, which is vectors and then check the similarity of the words in the job seekers qualification to the dictionary of words built for each job ID. This calculation will be for both sides. It is not a redundancy task. The preference from both sides will be add in the next layer. Any outliers will be detected and removed. Then the score of each job ID and job seeker ID will be sorted in a data frame in the DL. Equation (1) and (2) illustrates the computation of ScoreSeek for a particular JobID and ScoreJob for a particular CandidateID where:

- DicJob is the job description of dictionary words.
- DicSeek is the qualification dictionary word.
- SimJob is the similarity value calculated of one word from qualification against job description dictionary.
- SimSeek is the similarity value calculated of one word from the job description against the qualification dictionary.
- ScoreJob is the sum of SimJob.
- ScoreSeek is the sum of SimSeek.
- n number of words in the dictionary.

$$ScoreSeek_{jobID} = \sum_{1}^{n} SimSeek_{n} \quad (1)$$

$$ScoreJob_{candidateID} = \sum_{1}^{n} SimJob_{n} \quad (2)$$

3.3 Preferences layer

This layer will add the preferences, which is the weight of the word that is more important for both sides. The keywords or features will be displayed for both sides on the platform, and the user will sort his preference. Sorting the words will enable the model to give weight to each keyword. For example, if you are a recruiter in an academic field, you will give more weight on a candidate that has published a paper or if you want this candidate to be located in the same location where the job is, you will give more weight to the location. Then an in-depth text ranking will be applied, and the result will be stored back in the DL. Equation (3) and (4) illustrates the computation of ScoreSeek-pre for a particular JobID and ScoreJob-pre for a particular CandidateID where:

- SeekPre is the weight given by the job seeker to a word
- RecPre is the weight given by the recruiter to a word

$$ScoreSeek_{seekpre} = \sum_{1}^{n} SimSeek_{n} \cdot SeekPre \qquad (3)$$

$$ScoreJob_{recpre} = \sum_{1}^{n} SimJob_{n} \cdot RecPre \qquad (4)$$

4 Conclusion

This paper started with an overview of Saudi labor market and the government's tremendous efforts to improve it. We highlighted some of the current national projects, such as the initiation of Saudi Data and Artificial Intelligence Authority (SDAIA). The analysis concluded the need for a national central data repository and AI model that can think like a human in the recruiting field.

Then we presented the concept of a DL as a suitable data repository with significant advantages over traditional data repository. Furthermore, the paper proposed an AI recruiting model suitable for the Saudi labor market, which takes into consideration the preference of the recruiter and the job seeker and works to imitate the human brain. The AI model consists of three layers that work in sequence to match the job seeker with the best job ID. The first layer is called an Initial Screening layer. It builds groups of jobs from the same industry to gather and give a group ID by clustering them. The second layer is called a Mapping layer, which uses the RNN to find keywords, and Python NLTK library for word embedding Word2Vec to calculate the similarity of the keywords in the job seekers' qualification to the dictionary of words built for each job ID and vice versa. Then, the score of each job ID and job seeker ID will be sorted in a data frame in the DL. The third layer is the Preferences layer, which will add the preferences as a weight of the word that is more important for both sides. Then the result will be stored back in the DL. Further work of this paper is to implement this model and check its performance.

Future work will use NLP for the Arabic language to make it easy if the job description and the job seeker qualification were written in Arabic.

5 References

1. I. Nikolaou and J. K. Oostro, Employee recruitment, selection, and assessment : contemporary issues for theory and practice, Hove, East Sussex: Psychology Press, 2015.
2. J. A. Breaugh, "Recruiting and Attracting Talent," SHRM Foundation, United States of America., 2009.
3. W. Kenton, "Okun's Law," Investopedia, 26 Mar 2020. [Online]. Available: https://www.investopedia.com/terms/o/okunslaw.asp. [Accessed 07 Dec 2020].

4. Saudi vision, "GOVERNANCE MODEL FOR ACHIEVING SAUDI ARABIA'S VISION 2030," vision2030, 12 12 2019. [Online]. Available: https://www.vision2030.gov.sa/ar/node. [Accessed 7 12 2020] (in Arabic).
5. h. alsabeeh, a. aljassim, m. ahmed and f. hagemann, "Labor Market Dynamics in the GCC States," OxGAPS Oxford Gulf & Arabian Peninsula Studies Forum, Oxford, 2015.
6. K. Gill, E. Scott and L. Ward, "Understanding labour market information," Produced by the Department for Education and Skills, p. 5, 2004.
7. R. G. Ehrenberg and R. S. Smith, "Modern Labor Economics," PEARSON, p. 2, 2012.
8. A. Albaker and A. Alabdani, "Labor market chalenges in KSA," Saudi Arabian Monetary Agency, 2018 (in Arabic).
9. I. A. Wowczko, "Skills and Vacancy Analysis with Data Mining Techniques," open access informatics No. 2, pp. pp. 31-49, 2015.
10. S. A. Al-Zughaibi, "The importance of labor market reforms in restructuring the Saudi economy," 5 5 2014. [Online]. Available: https://www.alriyadh.com/915519. [Accessed 7 12 2020] (in Arabic).
11. A. M. Greenwood, "International definitions and prospects of underemployment statistics," Proceedings for the Seminario sobre Subempleo, pp. 8-12, 1999.
12. H. Stephen, "Is it possible to revive the labor market?," King Faisal Center for Research and Islamic Studies, Riyadh, 2018.
13. A. H. ,. Al Omari, "Characteristics of the labor market in Saudi Arabia," Economic writer Abdul Hamid Al-Omari, 30 09 2003. [Online]. Available: http://abdulhamid.net/archives/3450 [Accessed 7 12 2020] (in Arabic).
14. M. A. Al-Sudairy, "Launching the National Labor Observatory Portal to stimulate Emiratization and the organization of the labor market," Saudi press agency, 31 1 2019. [Online]. Available: https://www.spa.gov.sa/1880625. [Accessed 7 12 2020] (in Arabic).
15. G. O. Statistics, "Labor market 3Q 2018," General Organization for Statistics Labor force statistics and social conditions, Riyadh, 2018(in Arabic).
16. F. Altorki and R. AlEshakh, "Future features of the Saudi labor market," Jadwa investment, Riyadh, 2015(in Arabic).
17. G. a. f. statistics, "Genral authority for statistics Labor market statistics second quarter of 2020," Genral authority for statistics, Riyadh, 2020(in Arabic).
18. M. alsharif, "Specialists: 15 Percentage of fake Emiratisation in the private sector," 12 9 2018. [Online]. Available: https://tinyurl.com/y2rysjud. [Accessed 7 12 2020] (in Arabic).
19. P. Sundsøy, J. Bjelland and B.-A. Reme, "Towards Real-Time Prediction of Unemployment and Profession," Telenor Group Research, Fornebu, Norway, 2019.
20. SDAIA, "About SDAIA," SDAIA, 1 3 2019 . [Online]. Available: (in Arabic). [Accessed 7 12 2020] (in Arabic).
21. A. Zomay and S. Sakr, "Encyclopedia of Big Data Technologies," 01 June 2018. [Online]. Available: https://link.springer.com/referenceworkentry/10.1007%2F978-3-319-63962-8_7-1#howtocite.
22. H. Fang, "Managing Data Lakes in Big Data Era: What's a data lake and why has it became popular in data management ecosystem," in The 5th Annual IEEE International Conference on Cyber Technology in Automation, Shenyang, China., 2015.
23. P.P. Khine and Z. . S. Wang, "Data Lake: A New Ideology in Big Data Era," researchgate, 15 12 2017.
24. . A. Tolstoy and N. Miloslavskaya, "Big Data, Fast Data and Data Lake Concepts," Procedia Computer Science, p. 302, 2016.
25. M.R. Llave, "Data lakes in business intelligence: reporting from the trenches," Procedia Computer Science, p. 516–524, 2018.

26. P. Patel and A. Diaz, "Data Lake Governance Best Practices," 25 Apr 17. [Online]. Available: https://dzone.com/articles/data-lake-governance-best-practices.
27. B. Stein and A. Morrison, "The enterprise data lake: Better integration and deeper analytics," Technology Forecast: Rethinking integration, 2014.
28. b. satpute, "Enterprise Data Lake: Architecture Using Big Data Technologies - Bhushan Satpute, Solution Architect," 28 mar 2016. [Online]. Available: https://www.youtube.com/watch?v=hsq4s_l9ZDM. [Accessed 7 12 2020].
29. EDUCBA, "Data Lake Architecture," EDUCBA, [Online]. Available: https://www.educba.com/data-lake-architecture/. [Accessed 7 12 2020].
30. educba, "Data Lake Architecture," 3 2 2020. [Online]. Available: https://www.educba.com/data-lake-architecture/.
31. marionoioso, "A Big Data architecture in data processing," 22 8 2019. [Online]. Available: https://marionoioso.com/2019/08/22/a-big-data-architecture-in-data-processing/.
32. J. Brownlee, "10 Clustering Algorithms With Python," Machine Learning Mastery , 20 08 2020. [Online]. Available: https://machinelearningmastery.com/clustering-algorithms-with-python/. [Accessed 7 12 2020].
33. T. Zhang , R. Ramakrishnan and . M. Livny, "BIRCH: An Efficient Data Clustering Method for Very Large Databases," International Conference on Management of Data, vol. 25.2, p. 103, 1996.
34. A. Chopra, ,. Prashar and C. Sain, "Natural Language Processing," INTERNATIONAL JOURNAL OF TECHNOLOGY ENHANCEMENTS AND EMERGING ENGINEERING RESEARCH, vol. 1, no. 4, pp. 131-134, 2013.
35. T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient Estimation of Word Representations in Vector Space," arXiv, vol. 3, no. 1301.3781, 2013.
36. L. BORGES, B. MARTINS and P. CALADO, "Combining Similarity Features and Deep Representation Learning for Stance Detection in the Context of Checking Fake News," ACM Journal of Data and Information Quality, vol. 9, no. 4, p. 39, 2019.
37. I. Sutskever, J. Martens and G. Hinton, "Generating Text with Recurrent Neural Networks," Proceedings of the 28th International Conference on Machine Learning, pp. pp.1017-1014, 2011.
38. V. Ha-Thuc, G. Venkataraman, M. Rodriguez, S. Sinha, S. Sundaram and L. Guo, "Personalized Expertise Search at LinkedIn," arXiv, vol. arXiv:1602.04572v1 , no. [cs.IR], 2016.
39. K. JARVELIN and J. K. AL¨ AINEN, "Cumulated Gain-based Evaluation of IR Techniques.," ACM Trans. on Information Systems (TOIS), vol. 20, no. 4, pp. 422-446, 2002.
40. S. Cem Geyik, . Q. Guo, B. Hu, C. Ozcaglar, K. Thakkar, X. Wu and K. Kenthapadi, "Talent Search and Recommendation Systems at LinkedIn: Practical Challenges and Lessons Learned," arXiv, vol. arXiv:1809.06481v1 [cs.AI], 2018.