

Aleisa, Monirah Ali; Beloff, Natalia; White, Martin

Article

Implementing AIRM: A new AI recruiting model for the Saudi Arabia labour market

Journal of Innovation and Entrepreneurship

Provided in Cooperation with:

Springer Nature

Suggested Citation: Aleisa, Monirah Ali; Beloff, Natalia; White, Martin (2023) : Implementing AIRM: A new AI recruiting model for the Saudi Arabia labour market, Journal of Innovation and Entrepreneurship, ISSN 2192-5372, Springer, Heidelberg, Vol. 12, Iss. 1, pp. 1-41, <https://doi.org/10.1186/s13731-023-00324-w>

This Version is available at:

<https://hdl.handle.net/10419/290242>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.



<https://creativecommons.org/licenses/by/4.0/>

RESEARCH

Open Access



Implementing AIRM: a new AI recruiting model for the Saudi Arabia labour market

Monirah Ali Aleisa^{1*} , Natalia Beloff¹ and Martin White¹

*Correspondence:
Ma989@sussex.ac.uk

¹ University of Sussex, Falmer,
Brighton BN1 9RH, UK

Abstract

One of the goals of Saudi Vision 2030 is to keep the unemployment rate at the lowest level to empower the economy. Prior research has shown that an increase in unemployment has a negative effect on a country's Gross Domestic Product (GDP). This paper aims to utilise cutting-edge technology such as Data Lake (DL), Machine Learning (ML) and Artificial Intelligence (AI) to assist the Saudi labour market by matching job seekers with vacant positions. Currently, human experts carry out this process; however, this is time-consuming and labour-intensive. Moreover, in the Saudi labour market, this process does not use a cohesive data centre to monitor, integrate or analyse labour-market data, resulting in several inefficiencies, such as bias and latency. These inefficiencies arise from a lack of technologies and, more importantly, from having an open labour-market without a national data centre. This paper proposes a new AI Recruiting Model (AIRM) architecture that exploits DLs, ML and AI to rapidly and efficiently match job seekers to vacant positions in the Saudi labour market. A Minimum Viable Product (MVP) is employed to test the proposed AIRM architecture using a labour market dataset simulation corpus for training purposes; the architecture is further evaluated against three research collaborators who are all professionals in Human Resources (HR). As this research is data-driven in nature, it requires collaboration from domain experts. The first layer of the AIRM architecture uses balanced iterative reducing and clustering using hierarchies (BIRCH) as a clustering algorithm for the initial screening layer. The mapping layer uses sentence transformers with a robustly optimised BERT pre-training approach (RoBERTa) as the base model, and ranking is carried out using the Facebook AI Similarity Search (FAISS). Finally, the preferences layer takes the user's preferences as a list and sorts the results using the pre-trained cross-encoders model, considering the weight of the more important words. This new AIRM has yielded favourable outcomes: This research considered accepting an AIRM selection ratified by at least one HR expert to account for the subjective character of the selection process when exclusively handled by human HR experts. The research evaluated the AIRM using two metrics: accuracy and time. The AIRM had an overall matching accuracy of 84%, with at least one expert agreeing with the system's output. Furthermore, it completed the task in 2.4 min, whereas human experts took more than 6 days on average. Overall, the AIRM outperforms humans in task execution, making it useful in pre-selecting a group of applicants and positions. The

AIRM is not limited to government services. It can also help any commercial business that uses Big Data.

Keywords: Recruiting, Job seeker, AI, Natural Language Processing, Clustering algorithms, BIRCH, RoB-ERTa, FAISS, Data Lake, SBERT

Introduction

Artificial intelligence (AI) is one of the most rapidly evolving technologies today, and it has been used in a variety of application domains. Recently, AI technology has shone in the field of recruiting. Many researchers are working on expanding its capabilities with various applications that will aid in the recruitment process. However, importantly, having an open labour market without a cohesive data centre makes it difficult to monitor, integrate and analyse data to help reach the best match of a job candidate to a job vacancy.

Hiring and screening curriculum vitae for jobs is a time-consuming and labour-intensive process. Typically, each resume is written uniquely. Before hiring someone, recruiters must read many resumes and comprehend their content. Currently, human experts carry out this lengthy and laborious process. Moreover, in the Saudi labour market, this process does not use a cohesive data centre to monitor, integrate or analyse labour market data, resulting in inefficiencies such as bias and latency. These inefficiencies arise from a lack of technologies and, more importantly, from having an open labour market without a national labour market data centre. It is important to use labour market data to reduce the unemployment rate; this is the driving factor behind this study.

This paper proposes a model that will support the labour market from both a technical and database strength standpoint, to tackle the unemployment problem that has a negative effect on the country's gross domestic product. According to Saudi Arabia's General Authority for Statistics (GSTAT), the Saudi unemployment rate in the third quarter of 2021 was 11.3% (General Authority for Statistics, 2021). Furthermore, Okun's Law, states that a 1% increase in unemployment results in a 2% decrease in the Gross Domestic Product (GDP) (Kenton, 2020). Okun's. Consequently, Saudi Arabia's unemployment rate 11.3% is quite high for a rich and developing country with approximately 35 million people, 70% of whom are 'youths' (aged 20–55); 'youth' here means of an age eligible for work. Hence, Saudi Arabia's GDP may increase by approximately 22.6% if it has nearly zero unemployment. Considering a more realistic goal of reaching a 5% unemployment rate, which would increase the GDP by 12.6%, the significance of assisting in the reduction of unemployment is evident here.

Saudi Vision 2030 is a plan 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 organised by the Council of Economic and Development Affairs. Furthermore, it is accomplished in collaboration with the public, private, and non-profit sectors (Affairs_of_V, 2030, 2016). The Kingdom of Saudi Arabia is putting much effort into this area. However, there are numerous obstacles in the Saudi labour market, each with its own set of challenges. The significance of assisting in the reduction of unemployment is evident here. Previous studies in this domain, where AI serves in the recruiting field, has

concentrated solely on linear and non-linear models that, once trained, lack preference personalisation at the user level. Additionally, it served commercial applications more than governments services.

This paper is centred on the labour market in Saudi Arabia. In particular, it focuses on investigating how cutting-edge technology, such as AI and ML, can be exploited to process large national labour market datasets and store them more efficiently in repositories, such as DL, to match job seekers with job vacancies in the Saudi labour market. This study builds a new AI Recruiting Model for the Saudi Arabia Labour Market (AIRM). The AIRM can be used to effectively mimic how the human brain functions when selecting a job as a job seeker or as a recruiter looking for the best candidate to match a vacant position job, taking into consideration user preference. Since all resume data deal with large amounts of natural language as text data, this architecture employs the most recent new profound text semantic understanding technologies, such as contextual embedding, which has been used to capture complex query-document relations. This is achieved by retrieving similar candidates for a given job description and vice versa, and by identifying all the attributes or columns needed to fulfil the system requirements. Since millions of jobs are posted on different platforms leading to a tremendous amount of data, this study also addresses latency issues while retrieving similar job candidates in ample data space.

This process must be regularly altered to reflect the changes and new data in the labour market. By integrating the data from government agents’ databases, it will be possible to analyse all the data to predict the needs of the Saudi labour market and assist decision-makers. This study assumes that a platform system will be built and set up to be used by all sectors as a unified platform (front end) and that the proposed model that matches jobs with the candidates will function as a matching engine (back end); see Fig. 1

The final results will include recommendations for both recruiters and job seekers. Recruiters will be able to see the top ten recommended candidates, and job seekers will see the top ten recommended jobs based on several characteristics. A Minimum Viable Product (MVP) is employed to test the proposed AIRM architecture using a labour market dataset simulation corpus for training purposes, evaluated against three collaborating Human Resources (HR) professionals. As this study is data-driven in nature, it requires collaboration from domain experts.

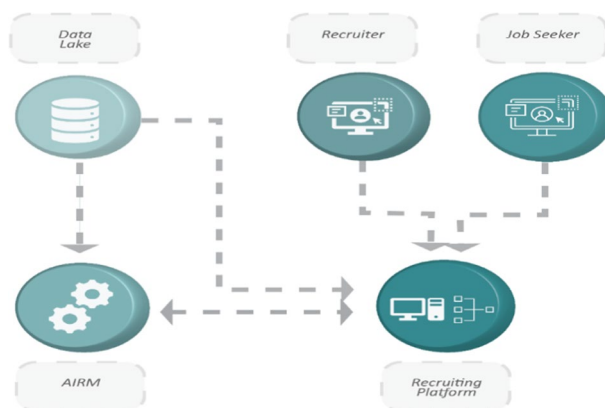


Fig. 1 Labour market platform connected to the AIRM

This MVP has yielded favourable outcomes. It considered accepting the MVP selection ratified by at least one HR expert to account for the subjective character of the selection process when exclusively handled by human HR experts. The research evaluated the MVP using two metrics: accuracy and time. The MVP had an overall matching accuracy of 84%, with at least one expert agreeing with the system's output. Furthermore, it completed the task in 2.4 min, whereas human experts took more than six days on average. Overall, the MVP of the AIRM outperforms humans in task execution, making it useful in pre-selecting a group of applicants and positions. The AIRM is not limited to government services; it can also help any commercial business that uses Big Data.

The AIRM will provide the government with a clear vision of outputs, inputs, labour market requirements and labour market behaviour; consequently, decision-makers may open new areas or fields of study in universities, thus making the local labour market simulate the international labour market. In addition, the AIRM will provide a comprehensive vision with the full cooperation of government agencies and the private sector, which is not currently available. Moreover, the AIRM can be generalised to run the model on different types of data sets, and implement the AIRM using any other language such as Arabic and compare it with the performance of this English version.

Background research

Inventions and innovations arise from a pressing need for them. The qualified workforce in the labour market is a focal point in Saudi Vision 2030 (Almaoasi, 2017). Saudi Vision 2030 aims to reduce the unemployment rate to 5% by 2030. The Saudi government spends a more significant proportion of its GDP on education as compared to the rest of the world (Harvard Kennedy School, 2021). Since the 1980s, Saudi Arabia has experienced rapid population growth, which has impacted its population structure; and a large number of youths are present in the community, increasing the demand for jobs. Ideally, employment is supported by talent, and talent changes the economy and affects the entire nation. Therefore, talent must be supported and developed if jobs are to be carried out by well-qualified citizens. Achieving a satisfactory level of unemployment could be accomplished by developing a data-driven matching engine. It would analyse data from various sources to map the relationship between job supply and demand in the labour market, which is too complex to accomplish manually. ML and AI techniques that can retrieve information about the current labour market situation and predict the required jobs will help all those involved with the educational system (parents, students and decision-makers) to know which skills are required in the market.

To build an optimal model that supports the labour market and to achieve the effective mapping of job descriptions to job seekers, the following steps are necessary:

- Define the business problem,
- Build a marketing database,
- Explore the data,
- Prepare the data for modelling,
- Build the model,
- Evaluate the model,

- Deploy the model and analyse the results.

Currently, Saudi Arabia has no DL or integrated system that links the two essential datasets, the education dataset and the full labour market dataset and other valuable data related to matching candidates with jobs. The integration of these datasets will play a significant role in bridging the massive gap in the Saudi labour market; moreover, using ML and AI techniques, vital knowledge can be extracted from the data. This process involves analysing data to find patterns and then extracting the information needed to support decisions. Unfortunately, these two crucial datasets are entirely separate. Consequently, a database that integrates labour market development would be ideal. The data required for development can come from education and labour market data and other essential databases that should not be overlooked when integrating data, such as data from the GSTAT, the Ministry of Labour and Social Development (MLSD), and the National Information Centre (NIC).

The Saudi government is working on the National Transformation Program to encourage agencies, especially government agencies, to develop electronic services for their audiences. This presents an excellent opportunity to encourage these agencies to integrate their databases, which will help with the search task. As a result, sufficient information about unemployed people will be available. For example, the MLSD, the Ministry of Civil Service (MCS), and the Human Resources Development Fund (HADAF) databases can be used. These data sources can be fed into the proposed AIRM, which will create as many records as possible in the database. All of these records will be analysed. Notably, the National Digital Transformation unit developed a 'New Road Map' for 2017 to 2019, emphasising the significance of national digital transformation (National Digital Transformation, 2020). This road map was recognised by developing an integrated innovative digital platforms and services system for the Ministry of Interior beneficiaries, the government sector, the private sector and individuals.

The NIC has a new digital strategy that enhances the centre's role in providing unique digital solutions and services to set up and operate intelligent digital platforms, thereby allowing its customers and partners to build and operate digital solutions and services using the latest technologies at lower costs and with greater efficiency and security. The NIC provides services to 13 cities, over 40 different agencies at the Ministry of Interior, and over 35 government entities with which the centre has data-exchange agreements (WAS_SPA, 2017). The cornerstone of this strategy aligns with the role that data centres play in achieving the desired national transformation in general and digital transformation in particular. Furthermore, the centre aims to strengthen the centralisation of and provide maximum protection to these data to extract the highest degree of benefit from the data at all levels.

Therefore, another approach emerged to classify job vacancies according to a target classification system. It can be used to build a language-independent knowledge base for analysis purposes instead of only matching CVs with job vacancies. This could be a new approach for natural language texts (Boselli et al., 2017). This approach will help with deep data analysis and enable data science to aggregate and group data to get meaningful information and build a dashboard for the decision-makers.

One study by Colombo et al. proposed developing a new set of tools for LMI by applying ML techniques to web vacancies and focused on the Italian labour market in particular. The tools designed for analysing skill needs in this study shed light on some key issues. Firstly, calculating different types of skills were required for each occupation; then, those skills were classified according to a standard classification, and then codes and measures of the relevance of digital skills were developed. Secondly, the correlation between soft and digital skills was shown, and the probability of automation of a given occupation. Thirdly, measures of the variation over time in the terminology used in describing occupations trades were developed (Colombo et al., 2018).

In another study done by Wowczko, trying to apply TF-IDF technique to utilise data mining techniques to map job descriptions with job seekers, is an outdated technique and does not represent words well in context (Wowczko, 2015). This study uses the latest technology in the field of natural language processing to solve the problem of correctly representing a word in the context of the sentence.

Literature review

This section introduces a comprehensive literature review to perform a thorough investigation of the Saudi labour market across multiple dimensions in order to immerse the reader in the subject.

The beginning of this literature review provides an overview of the labour market in the Kingdom of Saudi Arabia, focusing on statistical information. Following that, it explores the definitions of the labour market and the consequences of unemployment according to existing literature. The review then proceeds to compare the labour market efforts of Saudi Arabia with those of other countries. Moving forward, it discusses skills matching and identifies the challenges that impact the accuracy of training data when developing machine learning or artificial intelligence algorithms. Additionally, it defines machine learning (ML), artificial intelligence (AI), labour market intelligence, and text classification. Finally, the review addresses the data challenges specific to the labour market.

The Kingdom of Saudi Arabia is currently witnessing an unprecedented economic transformation, which has affected all government activities, aiming to create jobs and involve more Saudi women in the labour market. Another economic objective within the Saudi 2030 vision is to reduce foreign financial remittances (Albaker & Alabdani, 2018). The rapid revolution in the labour market might take an unexpected road, where changes in the skills needed for current jobs coincide with the emergence of new jobs, which requires an effective method of monitoring skills, which has not been available until now (Wowczko, 2015). The Saudi labour market relies heavily on foreign workers, especially in the private sector, due to two reasons. The first reason is the massive demand for workers in the oil sector and other industrial jobs. The second reason is the size of the Kingdom of Saudi Arabia, which needs large infrastructure projects which require temporary workers that work only for the duration of the project and therefore do not provide secure employment opportunities for Saudis (Albaker & Alabdani, 2018).

The labour market in Saudi Arabia is divided into a government sector that follows the Ministry of Civil Service (MCS), a semi-government sector and a private sector whose pension system is subject to the General Organization for Social Insurance

(GOSI). These three distinct labour markets have different characteristics. According to Vision 2030, the private sector will create most new jobs, with 4.5 million new work opportunities for Saudi women and men in the private sector by 2025. It also aims to boost Saudis' inclusion in the private sector by 50% by 2025. Given the current state of circumstances, this is a high lofty ambition (Privatization_Program, 2018).

The literature concentrates on the private sector since it is an essential primary driver in the labour market. Moreover, the economy is undergoing enormous transition, demonstrating how challenging the integration of the Saudi labour market is. With the lack of a transparent monitoring system, the aim is to keep the unemployment rate at its lowest level and employment at its highest level, which means a stable economic situation. Any practical examination of the labour market must be based on data that makes problem identification and analysis easier.

Abdul Hamid Al Omari is an economic specialist working for a Saudi financial agency and one of the well-known economic writers in Saudi Arabia; he summarised the most essential characteristics of the labour market in Saudi Arabia, which were developed by the Labour Force Council, as follows:

- The lack of adequate data on the Saudi labour market, employment and unemployment.
- The largest age group in the Saudi population is children and adolescents. This will be reflected in the increase in the working-age population in the coming years; the government need to plan for that group wisely.
- Saudi women's contribution to the labour market is low, just 35%. Although a large proportion of Saudi women applicants have university qualifications, there are limited opportunities available to Saudi women (General Authority for Statistics, 2018).
- 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.

A labour market can be defined as a mapping process that forms a mechanism to match demand with supply for employers and employees (Gill et al., 2015). Historical transformations in the labour market have generally been due to changes in labour conditions, and some factors have dramatically changed the characteristics and nature of the Saudi labour market in the past few decades, just as in other developed and developing countries. Several jobs have disappeared while new jobs have become available; and some are genuinely novel jobs that did not exist until a few years ago.

The labour market imbalance in the private sector is characterised by its heavy reliance on low-paid expatriate workers. This kind of employment cannot significantly contribute to knowledge transfer or empower the economy (Al-Zughaibi, 2014). Unemployment is one of the economic indicators of labour market performance, and it affects families by reducing their purchasing power, and the nation generally loses because of its impact on the economy. Unemployment is also a driver of migration patterns. The problems of poverty and unemployment have always been critical obstacles to economic development (Sundsøy et al., 2017). The main conclusion drawn about the impact of

labour market problems 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; and this study proposes a model to support the labour market from both a technical and database strength standpoint.

Restructuring the Saudi economy is a long-term strategic development objective, but it cannot be done in isolation by simply reforming the labour market, especially in the private sector, which relies heavily on low-waged and low-skilled expatriate labour. This excessive dependence on this type of employment reduces opportunities for the development of the Saudi economy. It also reduces the provision of job opportunities for Saudi citizens. Businesses that effectively use AI can cause a disturbing revolution with their new digital models and practices, allowing them to potentially transform the global economic business landscape. It appears that traditional attitudes frequently oppose the formation of a culture of novelty. A significant fulcrum implies a favourable degree of freedom in bringing possibilities and revelation to a new lens (Mishra & Tripathi, 2021). The radical solution is to obtain accurate data that can be integrated from all sources and analysed. In this respect, the Kingdom of Saudi Arabia government has made tremendous efforts towards digitalisation, as it has established an authority called the Saudi Data and Artificial Intelligence Authority (SDAIA).

The SDAIA was a new establishment in Saudi Arabia in 2019. It supports the Kingdom's Vision 2030, by promoting the Kingdom's digital capabilities and intention to build a data-based economy. The SDAIA works to regulate the data sector and enable innovation and creativity through its three arms: The National Data Management Office, the National Information Center, and the National Center for AI. It is 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 (SDAIA, 2021). The National Data Management Office in SDAIA is building a national data bank, which will regulate the data stream flowing from all government agencies. 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 (SDAIA, 2021).

One of the most well-known job portals is Europe's great online job portal (EURES), which is used in the Czech Republic, Denmark and Ireland. The use of the EURES portal for jobs is a novelty, and no other research has been found that is based on using publicly collected data from this source. Around twenty-six thousand employers have accounts that allow them to search for employees. EURES creates, updates and stores search profiles and can receive alerts by email. The excellent portal structure ensures a high degree of comparability across countries. This comparability comes from the secured vacancies and the accuracy of the uploaded data about vacancies uploaded by EURES employers. This comparability is an added value and a primary advantage of the EURES portal (Scarpetta & Sonnet, 2012).

Matching the skills that young people acquire at school and higher education and the skills needed in the labour market, is often a problem for the labour market; and it has become the biggest challenge to public and private sectors around the world as well for employers and job seekers. Economic and social policies must complement their efforts to solve the problem, and it is essential to focus on training efforts in growth strategies (Scarpetta & Sonnet, 2012). Three challenges will almost always affect the quality of the training data when building ML or AI algorithms. Firstly, if different employers use the same job title for the performance of different tasks, the same job title can be linked to two occupations, and this means that a complete classification must be at the individual level, not at the level of job titles. Secondly, this measurement challenge differs conceptually from the first in that a single individual performs functions that cross categories rather than two separate people with the same job title.

Employees who perform different functions may work two tasks but are only registered for one occupation; as a result, they fall at the margin of the occupation, which means that some job titles are marginal categories. Consider employees with the job title 'laboratory supervisor.' In many cases, these employees appeared to perform some tasks that would indicate they were assigned the occupation 'Research Facilitation Staff' and other tasks that would indicate they were assigned the occupation 'Research Staff'. Some laboratory supervisors, for example, serve as administrators for a university research lab while also conducting research within the lab. Thirdly, ambiguity is a measurement challenge, so an obscure title limits the ability to assign occupations to particular job titles. Uncertainty about these ambiguous titles would be a major cause of noise when dealing with ambiguous claims if manually classified occupations were subsequently used for training that could influence the learning process of ML algorithms. Ambiguous job titles are those such as 'administrative aide,' 'coordinator' and 'professional helper'. Some of these employees work in human resources, undergraduate admissions or a number of other offices that support basic university functions. Others may be actively interested in scientific research, either as a supporter or as a participant. To address title ambiguity, cooperation from data-submitting organisations is needed. These transitions potentially provide more leverage to specific job titles and also provide rich data about their career paths (Ikudo et al., 2018).

Traditionally, it has been argued that data are critical and play a significant role in connecting important government sectors—in this study case, education and the labour market. Data mining (DM) is simply a method for extracting knowledge from large databases based on data patterns to gain a competitive advantage. This method can also be used in HR to analyse, visualise and predict labour market trends. Organisations and governments can learn from current and potential employees' and citizens' behaviour to provide better services (Alsultanny, 2013).

The term 'labour market intelligence' (LMI) refers to frameworks containing algorithms to analyse the massive data in the labour market to support decision-making; for example, matching intelligently a job advertisement that includes two text fields and the title corresponding to the full description of a job vacancy (Boselli et al., 2017). Data extracted from unstructured texts focus on the E-recruitment process, which will support the decision-makers by matching candidate profiles with job descriptions by applying ML approaches. Occupations, trends and skills are labour market dynamics

that need to be deeply understood by both public and private labour market operators; and this is the added value of job vacancies on the web.

The importance of LMI comes from its ability to provide information about vocational and educational activities, available employees and the skills they have according to the needs of different sectors and regions in a much shorter time. For example, survey-based analyses require around a year to become available. Another example is that LMI can overcome linguistic boundaries, as instead of proprietary ones, standard classification systems can be used (Boselli et al., 2017). Little research exists about firm-level HR data on job titles with job classifications, although there are some intellectual foundations for occupational coding (Ikudo et al., 2018).

ML can be misused and confused with AI, as ML is a subset of AI, and according to the Oxford dictionary, it is "the capacity of a computer to learn from experience, i.e. to modify its processing based on newly acquired information" (Copeland, 2017; Larsson & Teigland, 2020; van der Zande et al., 2020). An ML definition from its primary field, the computational field, is: "ML is an evolving branch of computational algorithms designed to emulate human intelligence by learning from the surrounding environment. They are considered the workhorse in the new era of the so-called big data. Techniques based on ML have been applied successfully in diverse fields ranging from pattern recognition, computer vision, spacecraft engineering, finance, entertainment, and computational biology to biomedical and medical applications" (El Naqa, et al., 2015). All the above definitions suggest that machines can make decisions for humans by using an effective mathematical algorithm.

Many existing studies in the broader literature have shown that the most used ML model is Text Classification (TC) (Boselli et al., 2017). TC is the activity of labelling natural language texts from a pre-defined set, where the text classifier can learn by using an inductive process from a set of classified documents. It gives good results in categorising many real-life, web-based data such as social, news, media and sentiment analysis. TC can be used in the categorising of job openings. In most cases, CVs submitted via email or physical copy are more accurate than web-based ads. It is a quality concern that has an impact on the results.

Some of the worldwide job platforms, are the services and software solutions provided by JANZZ technology (Janzz, 2021). JANZZ is an employment service that enables recruiters to make significant progress in transparency, efficiency and accountability in skill and job matching, and the increasingly essential deployment of AI-based semantic technologies. This platform offers the most dependable, cutting-edge solutions for employment services to accurately match job searchers to relevant opportunities and generate the intelligent data needed to design and implement active labour market policies based on sound labour market intelligence. BurningGlass, Workday, TextKernel, and EmployInsight are examples of growing projects that leverage international standard taxonomies for skill matching created in recent years. Additionally, Google Job Search API has announced the classification of job vacancies using the Google ML service over the Occupational Information Network (ONET) (Boselli et al., 2017). Google aims to make searching for qualified job candidates easier by using AI techniques; and a private beta is available today for profile search (Boselli et al., 2017).

Arya is an online job search engine that works as an ML tool that finds sourcing patterns and draws potential candidates out of millions of online profiles. It gets smarter over time by using strategic feedback to learn from successes and failures (Arya Leoforce, 2021). At the 'Google's Cloud Next 2018' conference, they announced the 'Contact Center AI', with Google's Dialogflow package, aimed to leverage NLP to onboard customers efficiently. The Contact Center AI is available to more than ten enterprise vendors in all, with a popular platform like Twilio, Up-wire, and more than eight hundred customers signed up for alpha access (Wiggers, 2018).

A clustering method can be used for job vacancies, and similar job vacancies are clustered into one group to reduce the computational overhead involved in working with the actual data. New job vacancies will need this process to be done again, making it inefficient and requiring an online ML algorithm. The Stochastic Gradient Ascent (SGA) algorithm can be used for this reason. It is implemented to fulfil the requirements considering the new data point and automatically adjust the current analysis' statistics, setting weights to the features, adjusting the weights and generating a new set of suggestions with a new ranking. The closeness of the newly created vacancies is computed against the old centroids once, and the clusters of job vacancies are built, then the new vacancy will be added to the closest group, from which it has the shortest distance from the centroid.

This approach will help avoid high computation arising by re-computing the new clusters when a new job vacancy appears. However, the downside of this approach is that a newly created vacancy that is too far from every cluster might be assigned to a cluster that might not be optimal for it due to its distance. Determining a threshold can be used to address this issue, to determine if the distances computed between the new job vacancy and the centroids of every cluster are sufficient to assign the new vacancy into the existing clusters or to build another new cluster. For example, if the shortest distance between the centroids of clusters and the new vacancy is greater than the threshold, the new vacancy is considered an outlier and a new cluster are created, and the new vacancy is assigned to it. The structure of the Ejobz vacancy document reduces the computation every time a new job vacancy becomes available and creates new clusters to handle the new jobs, considering the value of the threshold (Chala, 2018).

There is a wide choice of platforms for e-recruiting. This research will focus on the data architecture and algorithm model and not extend to building a platform. What is important from the above review of best practices is to focus on their way of processing data. The idea behind job mapping is to document similarity analysis and clustering. Starting with the intellectual foundations for occupations:

1. Define all the occupations.
2. Translate concepts to standardised protocols.
3. Infer occupations from the information at hand.
4. Implement classifications for large data, given limited resources (Ikudo et al., 2018).

Firstly, set the document-to-document vector by a statistical approach after stop words are removed. The document vector is built from the essential words, and the words' importance is weighted. The weight is built up according to the word popularity

found. The outcome prioritises the relatively rare terms in the data set, and it is essential to take account of any synonyms. Vocabularies such as occupational names are used to guide the formation of the document vector to build and maintain a suitable vocabulary for a specific subject and improve the system's ability to distinguish between various job descriptions. Secondly, similarity analysis compares the vectors of documents using a range of statistical approaches, such as Term Frequency–Inverse Document Frequency (TF/IDF), Jaccard similarity, Cosine similarity, or/and Latent Semantic Analysis (LSA) techniques, also called Latent Semantic Indexing (LSI), and using the approach of vector space model and Singular Value Decomposition (SVD). Then a matrix of terms-by-documents is built to be used in the later stages to perform SVD on the matrix and find singular values that represent job descriptions as concepts in the document.

Furthermore, the Bidirectional Matching method can be used to carry out various operations and text analysis linked to the extraction and matching of multiple files utilising indexes based on document similarities, and to filter documents that do not contain a specific text. After matching, the model generates a matched data result and stores it in a database for further use (Chala et al., 2016). One of the approaches used is to cluster then compare. Firstly, clustering techniques are used to collect many groups representing particular noncognitive skills to create a noncognitive skill index. Secondly, the Lasso model decides which of these indices are essential for predicting labour market outcomes, and the results are compared with alternative index constructions (Mareckova & Pohlmeier, 2017).

With Python as a simple and freely accessible open-source programming language, key terms identified within the job description attribute find the skills needed for the occupational categories within the given dataset. A primary objective of using Python is to make the vacancy and skills analysis reproducible, reliable and cost-free (Wowczko, 2015). A study by Chala (2018) conducted on online data focused on analysing the text data to build a model against a newly produced vacancy. A new job seeker was compared for similarity to address the issues in qualitative mismatches compared to quantitative mismatches. Obtaining job seekers and vacancies through web mining, improving user experience in data collection from the job seeker, supplementing job seeker accounts in social network data and enhancing the content of vacancies through occupational standards data helped achieve the best matching. The study applied Natural Language Processing (NLP), web mining and ML, focusing on matching skills and job titles to vacancies and vice versa (Chala, 2018).

All the above clustering and mapping techniques have made seminal contributions helpful in conducting this research. Neural networks and K-mean also can be used.

Prior research suggests two occupational standards codes: the International Standard Classification of Occupations (ISCO) and European Skills, Competences, Qualifications and Occupations (ESCO). ESCO is an extension of ISCO and specifies the ISCO occupation groups into three segments:

1. Occupations.
2. Skills and competencies.
3. Qualifications.

Data from variable sources have different challenges, such as inconsistent structure and private information. An extensive text analysis and an elimination technique are needed to integrate diverse data. Missing data occur due to structure mapping, and dealing with missing data can be done by using special treatment in the analysis phase. There is a de-duplication issue for job vacancies. The data scraping process from multiple sources from the Internet might produce a job vacancy that appears in two or more of these portals. This duplication happens before any data pre-processing step. Collected vacancies are compared against one another to determine the level of their similarity. There are several challenges to integration involved in this process, such as dates, the difficulties in pre-processing and integrating the vacancy data collected from online sources, the names of the occupations and other subject names and the format (Chala, 2018).

DLs are a hot area in data modernisation. The DL is a new concept that is becoming a part of the core data architecture in modern data architecture; and will have a profound impact on data architecture used by governments and companies. As O'Brien pointed out, the DL is simultaneously an architectural strategy and an architectural destination and can be defined as a centralised repository for various data workloads. One of the benefits of DL is the ability to service multiple data application types based on a centralised data repository, such as data discovery, data science and enterprise Business Intelligence (BI) architecture (O'Brien, 2017). Moreover, previous research in this domain, where AI serves in the recruiting field, has focused solely on linear and non-linear models that, once trained, lack user preference personalisation. Furthermore, it was more focused on commercial applications than on government services.

The literature review conducted an analysis of the labour market in Saudi Arabia, specifically examining the existing literature that defines the labour market and the ongoing efforts related to the Saudi labour market. The focus was on various aspects, including government funding and initiatives, national labour market programmes, education, skills, qualifications, and global labour market projects. Throughout the review, a significant gap was identified. The crucial issue lies in the absence of a national data centre that can connect all relevant entities. This absence hinders the effective monitoring, integration, and analysis of labour market data, leading to several inefficiencies, such as bias and latency. These inefficiencies primarily stem from the lack of appropriate technologies and, more importantly, the absence of a unified national data centre for the open labour market.

Research methodology

Research methodology is a comprehensive approach that addresses questions like: How was the data collected or generated? How was it analysed? How are the tests interpreted? (Reich, 1994; Wilkinson, 2002). Hall and Kibler (1985) argue that prescriptive methodological analysis is unsuitable for AI research; however, researchers should make their methodological perspectives explicit when publishing research results to be better understood.

Despite recent advancements, there are signs that the field of AI still lacks the direction that a clear explanation of an appropriate methodology could provide. As can be seen, there is considerable disagreement about what constitutes significant research

problems. For instance, Hall and Kibler considered the divergence of opinions gathered in a research survey on knowledge representation. There was no clear consensus in the survey on what knowledge should be represented or what representation entailed. This type of ambiguity can also be seen in selecting appropriate research methodologies (Hall & Kibler, 1985). This study is rather old, and subsequent research will be more precise about how to create AI research in a way that is more compelling in its outcomes.

Researchers in the field of AI use a variety of methodological approaches. These various methodologies stem from the unique requirements of AI projects and the project's lifecycle, where AI projects are more data-centric than programmatic coding and are implemented in iterative steps (Walch, 2020). When AI tools are used successfully, human developers' creative potential is multiplied (Barenkamp et al., 2020). From the above literature, it is clear that there is no standard methodology for AI research, both academic and industrial, and it depends on the type of problem and the solution required. Building a robust data-driven prediction model, clustering data, constructing good decision rules, and helpfully visualising high-dimensional data are all things that AI researchers have in common with data scientists. Therefore, those steps are all iterative tasks that require knowledge and experience to complete. Furthermore, there have been few attempts to develop general methodological guidelines describing the overall process (François, 2008). There are four different data mining guidelines proposals known these days, and each has its characteristics. These methodologies are Fayyad's methodology or Knowledge Discovery in Databases (KDD) (Fayyad, 1996), Cios methodology (Cios is professor at the University of Colorado in Denver), Sample Explore Modify Model Assess (SEMMA) methodology, and the cross-industry standard process for data mining methodology (CRISP-DM). A summary comparison of them can be found in Table 1.

Choosing a research methodology or framework depends on the problem being investigated, as different design tasks have very different characteristics. This study focuses on using AI in the recruiting field. The suitable research type is Design Science Methodology (DSM), with an MVP as a proof of the new proposed architecture that consists of a set of AI, ML and a data repository to fulfil the research objective. In this research, we used the NLP technique. NLP offers significant mining opportunities in free-form text,

Table 1 Differences between methodologies

Description	KDD	Cios	SEMMA	CRISP-DM
Objective determination	X	X		X
Data collection	X	X	X	X
Data cleansing	X	X	X	X
Data reduction	X		X	X
Problem reformulation	X			
Data exploration	X		X	
Tool selection	X		X	
Model construction	X	X	X	X
Model validation	X	X	X	X
Result interpretation	X	X	X	X
Deployment	X	X		X

particularly for automated annotation and indexing prior to text corpora classification. Limited parsing capabilities can significantly aid in determining what an article refers to. As a result, the spectrum from simple NLP to language understanding can be highly beneficial. NLP can also contribute significantly as an effective interface for stating hints to mining algorithms and visualising and explaining knowledge derived by a KDD system (Fayyad, 1996). When these methodologies are compared to the research objective, it is clear that Fayyad’s KDD is the best framework to use. The KDD process is iterative and interactive, with many steps and many decisions made by the user; see Fig. 2.

The following are the necessary steps of this methodology:

Business problems definition: (Business Objectives) to understand business objectives, then convert them into sub-problems to develop the research plan.

Data understanding: in other words, exploring data starts with an initial data collection (Exploratory Statistics, Data Visualisation), understanding data after collection and verifying data quality. Further trends and relationships between attributes will be observed via data visualisations.

Data preparation: generally called data cleaning and pre-processing. This step involves selecting the necessary data, cleaning it (imputation, duplicate, fuzzy matching, etc.), reformatting it to fit, and integrating it once the model is complete; then creating the model and choosing between model techniques or mixed models.

Model evaluation: using different evaluation methods to guarantee high accuracy. We evaluated the AIRM by using two metrics: first, the accuracy and then the time.

Deployment: creating an MVP form AIRM and providing users’ permissions, as well as presenting the findings to recruitment professionals to test the accuracy of the AIRM. Also to make others use it and give feedback to enhance the model.

AIRM research framework

The AIRM architecture was proposed in the AIRM framework in Aleisa et al. (2021) based on KDD. It is tailored to the specific problem the research addresses, as shown in Fig. 3.

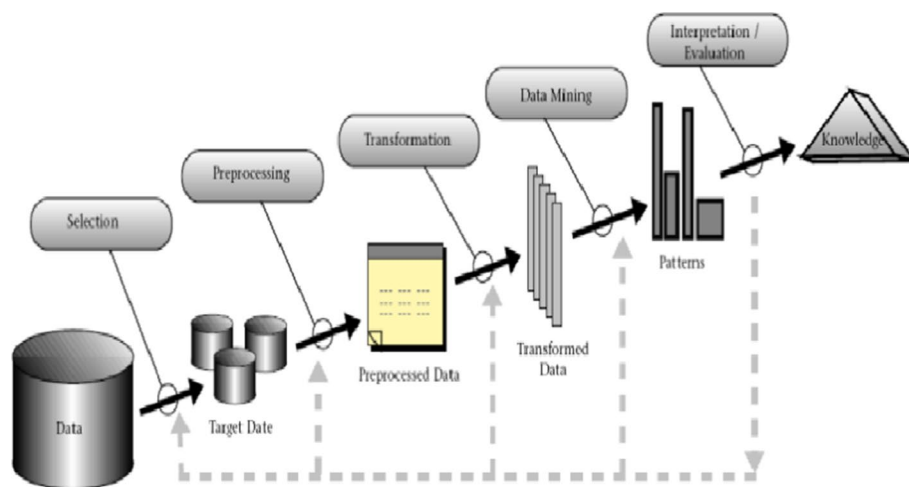


Fig. 2 Steps involved in a typical KDD process (Fayyad, 1996)

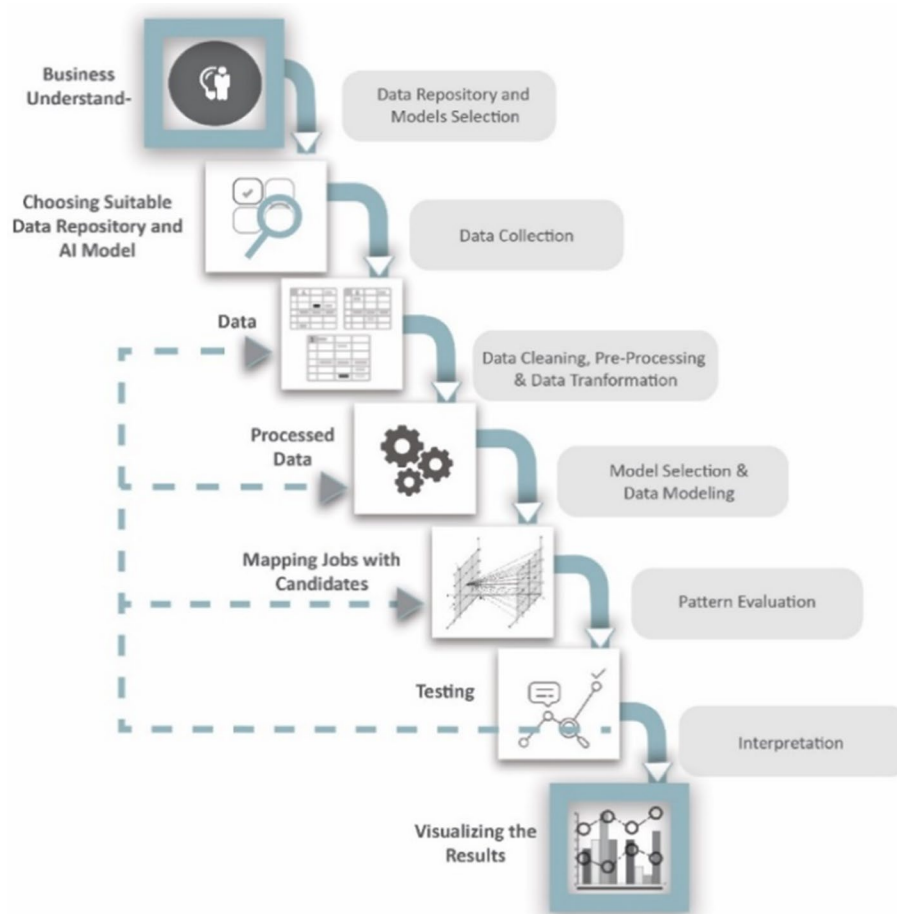


Fig. 3 The AIRM research framework

It starts with business understanding, data gathering, exploring, modelling, evaluation or interpretation and knowledge discovery. The framework will adopt an iterative approach. This research examines how AI technology can help enhance the Saudi labour-market by decreasing the gap between recruiters and job seekers; this is intended to empower the Saudi government's immediate and strategic decisions by providing comprehensive insights into the labour-market and expediting the recruiting process. An AIRM architecture is proposed to assist the labour-market by analysing the current Saudi labour-market. Accordingly, an approach with the following aspects is outlined: (1) a new data storage technology approach and (2) new ML and AI models, with three layers to extract the relevant information from the data of both recruiters and job seekers by exploiting ML and NLP for matching job candidates; for this, a suitable data repository technique with three processing layers, each with a different appropriate model, will be used.

Business understanding

The first step focuses on an in-depth analysis of the Saudi labour market in terms of the national projects that currently support the labour market, labour market growth,

the current recruiting system and the new Saudi data and AI authority. This analysis is carried out by reading and analysing all the topics related to the development of Saudi labour market projects. Furthermore, a thorough examination of the available data will enable better understanding of Saudi labour market requirements, allowing the examination to grow and develop productively. The overarching goal of this phase is to answer some intriguing research questions:

What are the specific challenges and gaps in data organisation for the labour market and recruitment services in Saudi Arabia?

What is the best technological approach for a repository for storing country-wide/national data?

Could an AI and Deep Learning approach provide efficient skill-to-need match-making recommendations using advanced natural language processing and big data techniques, specifically for the Saudi labour market?

What is the most suitable architectural framework and new recruitment model as a new approach to improve recruitment services (validated on the Saudi labour market case study)?

The proposed AIRM introduces a set of ML and AI models and a state-of-the-art data repository technique that will work with national data in response to the first three questions. The aim of the AIRM is to connect Saudi job seekers with the Saudi labour markets. To address the fourth question, AIRM proposes a framework with a data repository and three layers: an Initial Screening layer, a Mapping layer and a Preferences layer. The three layers collaborate to determine the best job ID for the job seeker.

Choosing a suitable data repository

It is essential to propose a suitable data repository to use after understanding the business problem and the type of data dealing with it. This research proposes a state-of-the-art data repository for reasons detailed in the proposed "[AIRM architecture](#)".

Choosing suitable models

This study implements a cutting-edge data repository as well as ML and AI models. It is critical to choose a suitable architecture that comprised an appropriate data repository, and ML and AI models, after understanding the business problem and the type of data at hand.

Table 2 Job and candidate attributes

Jobs data	Candidate data
Job title	Candidate personal information
Responsibility	Candidate qualifications
Qualification	Candidate achievements
Skills required	Candidate skills
Job location	Candidate preferred job location
Industry	Candidate preferred job industry
Job level	Candidate preferred job level

Data collection

At the start of the work on this research, the researcher did not find any actual Saudi labour market datasets that would be openly available for research purposes (e.g., for helping to train the AIRM). To mitigate this issue, we have created a simulation corpus for training purposes, which emulates a real labour market and assists in training the AI models. Attributes for this data set are shown in Table 2. In order to create the dataset, the data cloning process first collects data from available job advertising websites on the Internet. Some records were added manually to the dataset.

Once the dataset has been developed, job and candidate data then subsequently are cleaned before being fed into the AI model for training. This training corpus, which is openly available on Github, may be useful for other researchers to assist in similar simulations and AI model training tasks. Being an open source, it can also be improved/extended by future research.

Pre-processing data

This step involves searching for missing data and removing noisy, redundant and low-quality data from the data set to improve the data's quality and effectiveness. Specific algorithms are used for searching and eliminating unwanted data based on attributes related to the application. As part of data pre-processing, relevant features were selected from both job descriptions and candidate profiles. The pre-processing was carried out to ensure that the most relevant attributes were used. The job dataset's responsibilities, description and qualifications were chosen for further study. Similarly, experience, education, skills and certifications were chosen as relevant features from candidate profiles.

In order to ensure that data were free from noise, data cleaning was conducted as a second step. After exploratory data analysis, undesirable special characters were removed from the text. Numbers and stop words were not removed, considering that transformer models need raw input. Removing stop words can change the semantics of English sentences, and hence they were not excluded during data cleaning. The text was also lower cased as part of pre-processing step and to prepare the data to be fed to the algorithms. Hence, the data needed to be in consolidated and aggregated forms. The data were consolidated based on functions, attributes, features, etc.

Mapping jobs with candidates

This step of the AIRM research framework concerns the DL and three layers exploiting ML and AI to extract relevant information from the DL of both recruiters and job seekers to map them; see Fig. 4. The architecture uses ML and AI models, in particular:

1. For coarse clusters, BIRCH (Balanced Iterative Reducing and Clustering Using Hierarchies) clustering is used (Zhang et al., 1997). This ensures that search latency is kept as low as possible. Jobs and Candidates are categorised. There are, for example, a total of 20 clusters. If a candidate belongs to the fifth cluster, he or she will be able to find a job in that cluster as well. When looking for similar jobs, this eliminates latency (Janrao & Palivela, 2015).

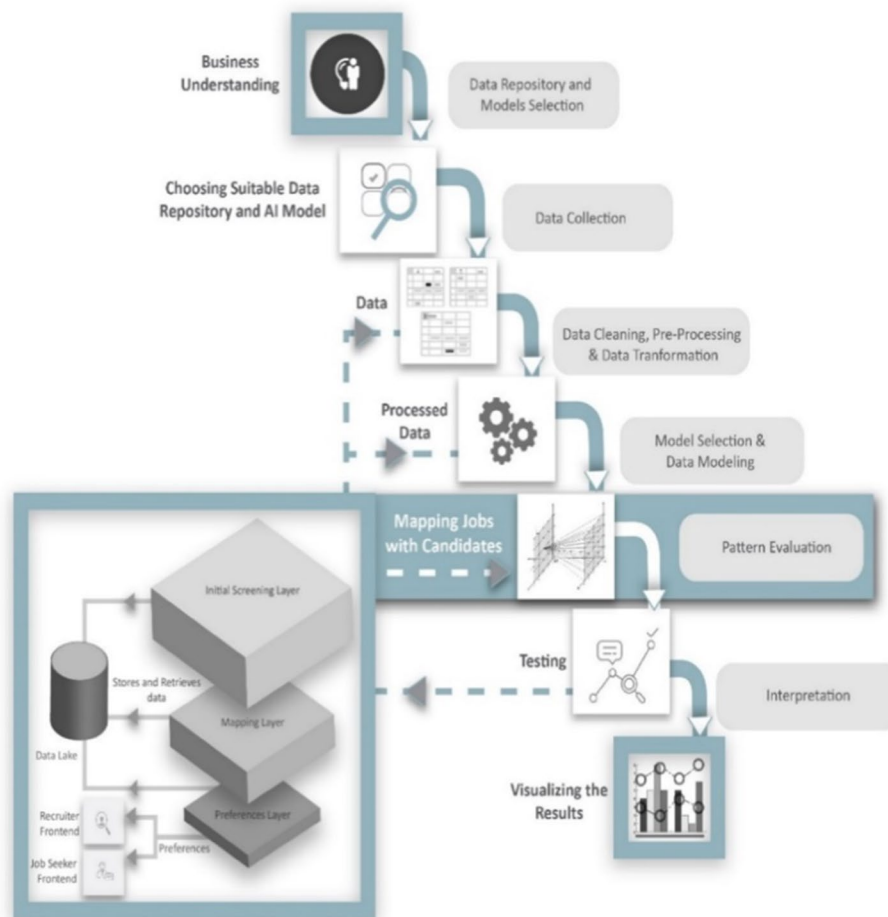


Fig. 4 The three layers in the AIRM research framework

2. Sentence transformers with RoBERTa as a base model are used (Reimers & Gurevych, 2020). This model has been trained on Natural Language Inference (NLI) and semantic textual similarity (STS) data. This transformer-based model is used to create embeddings for the job description and the candidate profile. The model is optimised to find similar documents because it is trained on a regression objective in a Siamese network fashion.
3. Once created, the embeddings are indexed in the Facebook AI Similarity Search (FAISS) (Johnson et al., 2019). FAISS is used to perform an approximate nearest neighbour search due to its extremely low latency. FAISS is compatible with both GPUs and CPUs. It can also be easily scaled. The goal is to use FAISS to find ten similar jobs for a candidate and vice versa. FAISS eliminates the computational overhead associated with traditional cosine similarity.
4. The user's preferences are then added using the pre-trained cross-encoders model (SBERTdocuments, 2021). To re-rank the result list, the weight of the more important words for users from both sides is considered.

Hence, the AIRM is a more non-linear method. It deals with text, language processing and deep learning. It is a tool that helps discover trends from a data set using AI models.

Testing

Since the AIRM uses a combination of unsupervised ML and AI algorithms, in order to test the performance, a group of external HR domain experts will check output samples and label the result as 'Good match' or 'Bad match'. Such feedback will help improve the AIRM's performance. Three human HR and recruitment specialists will work manually to review the AIRM findings for ten picked jobs, with the top ten comparable documents returned as candidates for each job. Human interaction is essential to evaluate the algorithm's performance for two reasons: first, we are utilising unsupervised algorithms, and second, selecting a candidate for a position is a subjective decision. For the assessment assignment, we will label the predictions as 'category three' if all three evaluators agreed on them, 'category two' if two evaluators agreed on them, and 'category one' if just one evaluator feels it matches; otherwise, we will label them as 'category zero'.

Visualising and discussing the results

Data visualisation is a critical step in ML and AI. Visualisation is essential when attempting to examine a dataset and extract some information in order to learn about it, as well as to detect patterns, corrupt data and outliers, to analyse massive amounts of information and make data-driven decisions. Data visualisation may be used to express and emphasise meaningful relationships in helpful plots and charts. These patterns can be deduced and presented in different plot forms, such as distribution plots, box plots, violin plots, line plots, bar plots and scatter plots, in which the preceding step's information is applied to the specific application or area in a visual manner. Dashboard structures can be built to visualise recruiting data that are endless. Data visuals can show inferential and descriptive statistics.

Repository techniques: Data Lake

Getting all the data in one place will support integrating it to allow data engineering to clean it. Then, data scientists can analyse it and apply ML algorithms to the data. This section provides a background for the DL and its utilisation.

James Dixon was the first to mention the concept of a 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 (Quix & Hai, 2018). 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" (Fang, 2015). Data are stored in the DL in their original format, whether structured, unstructured or multi-structured. Once data are placed in the lake, it is available for analysis (Khine & Wang, 2018). A comprehensive discussion about the DL, with its architecture build characteristics, data types, its use and how suitable it is for the AIRM, was presented in our previous work in (Aleisa et al., 2021). In this paper, we implemented the proposed architected and analysed the results.

Data Lake: Kylo

This section will outline Kylo, a new open-source DL, how powerful it is and how well it fits with the AIRM. Kylo is a high-performance DL platform built on Apache Hadoop

and Spark. It provides a turnkey, business-friendly DL solution, complete with self-service data ingest, preparation and discovery. It is a web application layer with features tailored to business users like data analysts, data stewards, data scientists, and IT operations personnel. Kylo incorporates industry’s best practices for metadata capture, security and data quality.

Moreover, Kylo offers a flexible data processing framework (based on Apache NiFi) to create batch or streaming pipeline templates and enable self-service features without jeopardising governance requirements. It was created by Think Big, a Teradata company, and is used by a dozen major corporations worldwide (Think Big, 2018). The Apache Software Foundation’s Apache NiFi project automates the flow of data between software systems, using the Extract, Transform and Load (ETL) paradigm (Apache NiFi Team, 2021). Think Big has reaped significant benefits from the open-source Hadoop ecosystem and has chosen to open-source Kylo to give back to the community and improve value; see Fig. 5.

In most cases, the workload is allocated to the cluster with the most processing power. Kylo orchestrates pipelines using Apache NiFi. With 180+ built-in processors, NiFi can connect to various sources and conduct lightweight conversions on edge. Kylo’s NiFi processor extensions can call Spark, Sqoop, Hive and even traditional ETL tools. SQL transformations are the focus of ETL solutions, which use their unique technology. The majority of data warehouse transformations are focused on importing normalised relational schemas like a star or snowflake. ELT tends to follow Hadoop data patterns (Think Big, 2018). Kylo’s added values are as follows:

- Kylo is a modern web application that runs on a Spark & Hadoop cluster’s Linux ‘edge node.’ Kylo includes a variety of custom routines for DL activities that use Spark and Apache Hive.
- Kylo’s scheduling and orchestration engine are Apache NiFi, which provides an integrated foundation for developing new 200-processor pipelines (data connectors

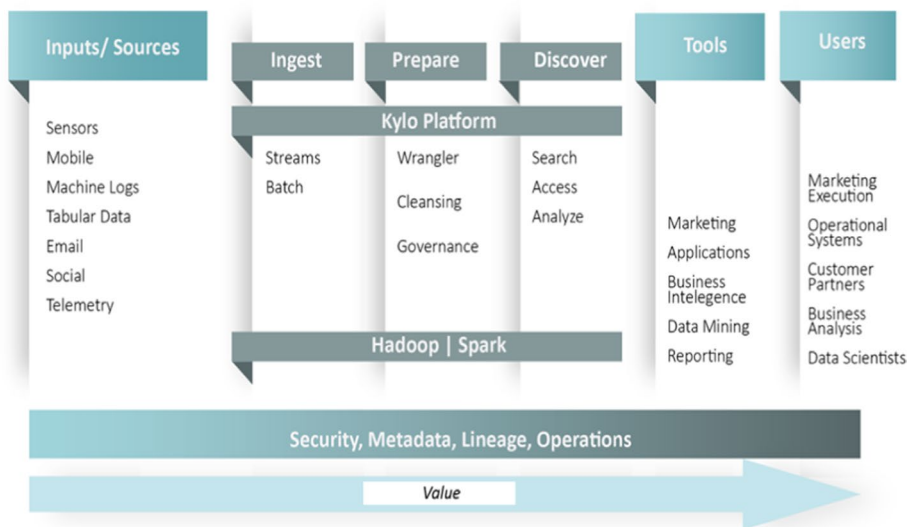


Fig. 5 Kylo architecture: adapted by the authors from Think Big (2018)

and transformers). Kylo includes a built-in metadata server that is currently compatible with MySQL and Postgres databases. For cluster monitoring, Kylo can connect to Apache Ranger or Sentry, as well as CDH Navigator or Ambari.

- ‘Write-once, use many times’, is one of Kylo’s additional values. Although NiFi is a robust IT tool for constructing pipelines, most DL feeds only use a few distinct flows or patterns. IT can develop and register a NiFi template as a data processing model for feeds using Kylo.
- Web modules provide important DL functionalities such as metadata search, data discovery, data wrangling, data browsing and event-based feed execution to connect flows.
- Data feeds can be monitored using the Operations Dashboard UI. It delivers feed health monitoring and related infrastructure services from a central location.

In order to implement the AIRM, we did not need expensive or complicated ETL technologies for Hadoop. We used Kylo, as it is more than enough for the AIRM. It uses Spark to wrangle and prepare visual data transformations with all the added values mentioned above over Apache NiFi.

Machine learning and artificial intelligence

This section emphasises a fundamental understanding of ML algorithms required to construct the AIRM, such as clustering and NLP algorithms. Before considering any AI algorithms, it is vital to prepare the data set. Some steps are required, such as scaling or normalising the data and imputing missing values. The feature values of each observation are represented as coordinates in n-dimensional space to calculate the distances between these coordinates. If these coordinates are not normalised, the results may be incorrect. It is also necessary to deal with missing, null or inf. There are several methods for dealing with such values, such as removing them or inputting them using mean, median, mode or advanced regression techniques (Rani & Rohil, 2013). We pre-prepared the data for the AIRM to ensure optimal results.

Clustering algorithm

For the reader’s convenience, this section will highlight some clustering algorithms to show the purpose of selecting a particular clustering algorithm. Clustering is the process of discovering a series of trends in a set of data; and it has produced positive results, especially in classifiers and predictors (Homenda & Pedrycz, 2018). Entities in one cluster should be dissimilar to entities in another cluster as much as possible. Clustering is an unsupervised approach that can spot erroneous class names, outliers and mistakes (Frades & Matthiesen, 2010). There are two types of clustering: hard clustering and soft clustering (SHARMA, 2019). There are several methods for measuring the distance between clusters to determine clustering rules, commonly referred to as Linkage Methods (Rani & Rohil, 2013). K-means or partition-based algorithms (Aggarwal & Reddy, 2014), hierarchy-based algorithms, fuzzy theory-based algorithms, grid-based algorithms and density-based algorithms are the five popular clustering algorithms (Aggarwal & Reddy, 2014). K-means is the most popular of all clustering algorithms.

BIRCH (Balanced Iterative Reducing and Clustering Using Hierarchies) is a clustering algorithm that excels at clustering extensive datasets. BIRCH handles large datasets by producing a more compact summary that retains as much distribution information as possible before clustering the data summary rather than the original dataset. The cost of BIRCH I/O is proportional to the size of the dataset: a single scan yields good clustering, and one or more additional passes can optionally be used to improve the quality even more. By evaluating its running time, memory usage, clustering quality, stability and scalability and comparing it to other existing algorithms, we argue that BIRCH is the best available clustering method for handling enormous datasets. Its architecture also supports parallel and concurrent clustering, and performance can be tuned interactively and dynamically based on data knowledge gained (Zhang et al., 2016). The AIRM will focus on BIRCH. The similarity between the clusters is calculated from dissimilarity measures like the Euclidean distance between two clusters. So, the more significant the distance between two clusters, the better the result (Pathak, 2018) BIRCH is suitable for the AIRM because of the four reasons mentioned in this paper in Sect. “[Justification of the implemented solution](#)”.

Natural language processing

In this paper, the AIRM will use two State of the Art (SoA) tasks: text similarity for the similar candidate to job matching and vice versa, and fast retrieval of similar documents. Text similarity refers to the comparison of vectors in hyperdimensional space. These vectors are of two types: sparse and dense vectors. Sparse vectors deal with Bag of Words (BoW) and TF-IDF (Term Frequency-Inverse Document Frequency) (Qaiser & Ali, 2018). There are few drawbacks to the TF-IDF approach. Since it is a statistical method, it fails to account for the order of words. Another drawback with such an approach is that it does not consider the semantics of a text. These problems of statistical methods can be overcome by other algorithms such as Word2Vec (Mikolov et al., 2013). Word2vec outperforms statistical methods such as BoW and n-gram models on various tasks. It also overcomes the affliction of dimensionality associated with statistical methods like TF-IDF and BoW. Word2vec work with a feed-forward neural network with a language modelling task and optimisation techniques such as Stochastic gradient descent. Word2vec has some drawbacks as it is not able to handle words with different contexts; and has the same representation for words with different contexts. For instance, if we consider two sentences, *I want to open a bank account* and *I want to sit near a river bank*. Here, word *bank* has different meanings in both sentences and the Word2vec model is not able to capture this contextual information. Also, Word2vec does not provide sentence embeddings. For sentence embeddings, word vectors need to be averaged. All these drawbacks can be well-handled by transformer-based models.

Transformer models were introduced in 2017; and the NLP field was revolutionised when Google introduced BERT (Devlin et al., 2019).

Transformer models like BERT achieve SoA results on the majority of NLP tasks. The transformer is a model architecture that completely avoids recurrence, favouring drawing global dependencies between input and output. It can be trained significantly faster than architectures based on recurrent or convolutional layers (Vaswani et al.,

2017). They even outperform sequence-to-sequence models like RNN (Recurrent Neural Networks). RNN-based models such as BiLSTM and GRU read input data sequentially. Another problem associated with RNN-based models is that their performance degrades as the input sequence length increases. Transformer-based models use an attention mechanism where each data is read once as a sequence of words. BERT is pre-trained on a massive corpus. Pre-training of such models require extensive data as well as substantial computation resources. Pre-training of BERT consist of Masked Language Modelling (MLM) and Next Sentence Prediction (Devlin et al., 2019). Other algorithms can overcome BERT problems like RoBERTa.

The transformer-based RoBERTa model was introduced in 2019 and outperformed BERT in many NLP tasks (Liu et al., 2019). RoBERTa has the same architecture as that of the BERT model. However, there are some differences in training the RoBERTa model. RoBERTa is pre-trained with larger batches and on a more extensive corpus. As compared to BERT, the RoBERTa model does not use Next Sentence Prediction (NSP) for the pre-training step (Liu et al., 2019). If NSP is not there, we do not need additional pre-training steps, and the model converges faster with better results. Due to its better performance, RoBERTa was chosen over BERT for generating embeddings. The RoBERTa base has 12 encoders stacked over each other compared to RoBERTa large, which has 24 encoders.

Figure 6 shows the architecture of an encoder in the RoBERTa model, which consists of self-attention and feed-forward forward layers.

For the re-ranking process, MS MARCO cross-encoders is used. MS MARCO is a Machine Reading Comprehension dataset on a vast scale built by Microsoft AI & Research (Bajaj et al., 2018). MS MARCO works as a large-scale information retrieval corpus created with the help of the Bing search engine and real-world user search queries. The cross-encoder model trained on MS MACRO can be used for semantic search; the model will find passages relevant to the search query. The training set contains about 500k samples, while the corpus has over 8.8 million paragraphs (SBERTdocuments, 2021). This re-ranking technique is used in the third layer of the AIRM architecture.

AIRM architecture

The proposed solution architecture AIRM is the topic of this section. It was constructed to create a proof of concept or a minimal viable product (MVP). The system extracts valuable information from the existing DL using data from both sides, recruiters and job seekers. The Initial Screening layer, the Mapping layer, and the Preferences layer make

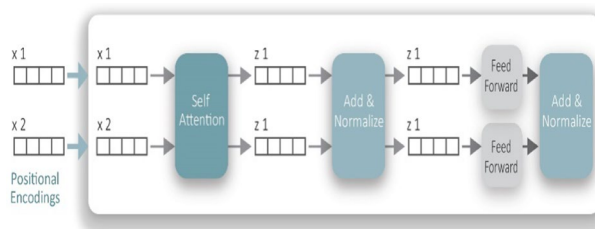


Fig. 6 RoBERTa encoder architecture adapted by the authors from Vaswani et al. (2017)

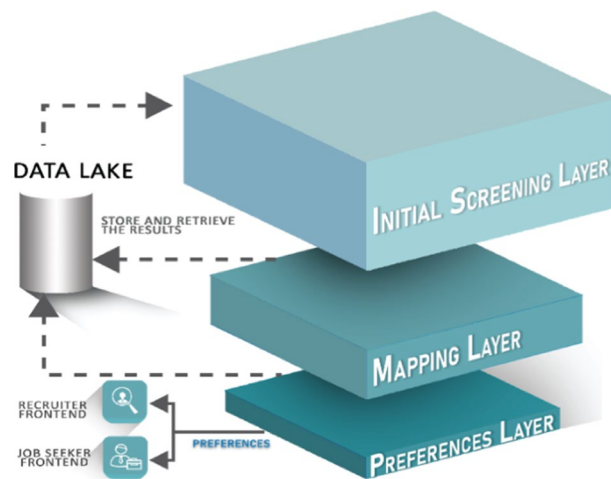


Fig. 7 Proposed AI recruiting model (AIRM)

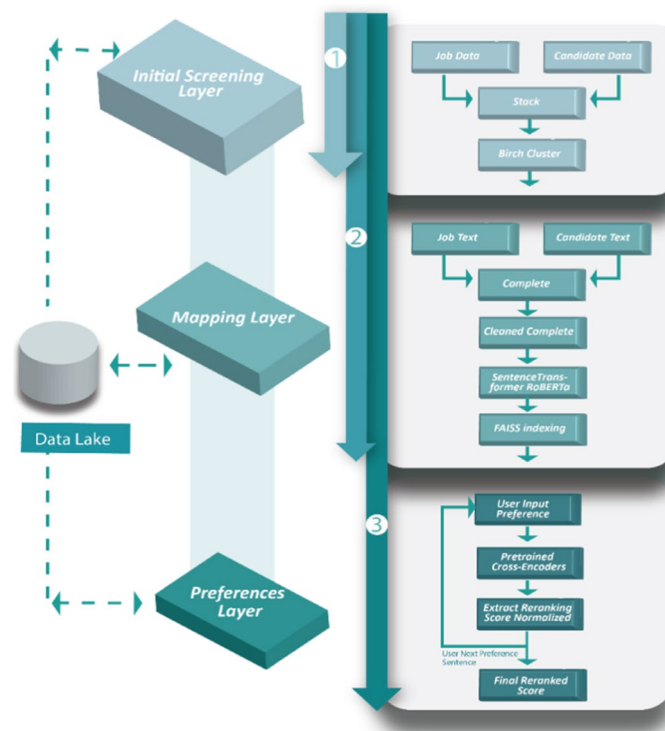


Fig. 8 The process steps in each of the AIRM's layers

up the AIRM architecture. The three layers work in sequence to match the job seeker with the best job ID. It retrieves and sends data from/to the DL; see Fig. 7.

The AIRM model has the following requirements:

- A national recruiting platform already exists.
- All recruiters' and candidates' data are clean, prepared for analysis and stored in the DL.

- Candidates can be tagged to one job or more, and the job ID can be tagged to one candidate or more.
- A comprehensive directory of job specialities in the DL, with instant updating.

The following three sections will focus on elaborating on code in each layer and dig deep into the process steps in each of the AIRM's layers, as shown in Fig. 8.

Data Lake

Getting all the data in one place will support integrating it to allow data engineering to clean it. Then, data scientists can analyse it and apply AI and ML algorithms to the data. The DL manages raw data as it is ingested from multiple data sources. It does not require cleansed or structured data, as mentioned in Sect. 2.2. A DL is a novel approach that makes use of big data technology. The DL saves data in its original format, whether structured, unstructured, or multi-structured. Once data are placed in the lake, it can be analysed. A comprehensive discussion about DL, with its architecture, build characteristics, data types, use, and how suitable it is for the AIRM, was presented in our previous work in (Aleisa et al., 2021). The benefit of the DL is that it is a hybrid data management solution that can handle big data issues while also enabling new levels of real-time analytics. They are highly scalable systems that can handle huge data volumes and receive data in its native format from various data sources on low-cost technologies that improve raw data capture, refinement, archival and exploration within an enterprise.

Initial screening layer

This layer will work as a preparation phase. It will use the BIRCH cluster algorithm to build cluster groups of job specialisations, which will enable the second layer to treat each cluster speciality separately. This layer's input is from both sides, the recruiters' data and job seekers' data. From the recruiters' data, the industry name, job level, job location and whether the employment period is full-time or part-time is required; data from the job seekers' side should include industry, location, employment period and whether it is full-time or part-time. 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 AIRM model's computational requirements for the next layer to enable the Mapping layer to work only with the needed group ID. Hierarchical clustering is a common form of unsupervised learning. It is suitable for this new proposed model, compared with other clustering algorithms. The similarity between the clusters is calculated from the dissimilarity measures like the Euclidean distance between two clusters. So, the larger the distance between two clusters, the better (Pathak, 2018). The data set must first be prepared before the data clustering process can begin. Scaling or normalising the data and missing value imputation are necessary steps in this layer. The feature values of each observation are represented as coordinates in n -dimensional space (n is the number of features), and the distances between these coordinates are calculated to normalise the data. Figure 9 shows the workflow in this layer.

The first task in this layer is to stack the two data frames over each other. The second task is using TF-IDF vectorisation to convert the text to numeric representation. Here,

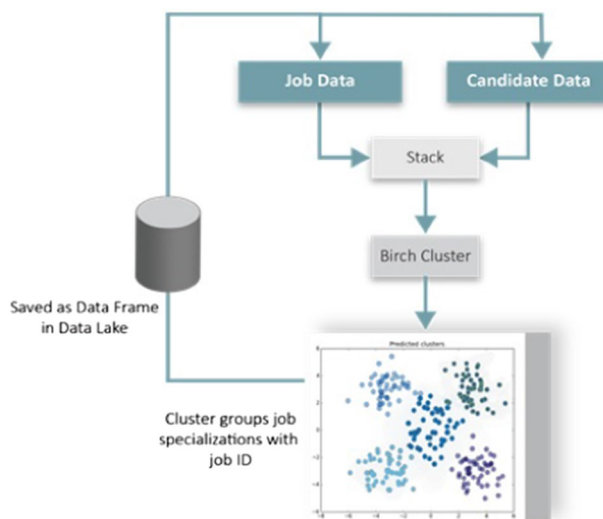


Fig. 9 Tasks in the initial screening layer

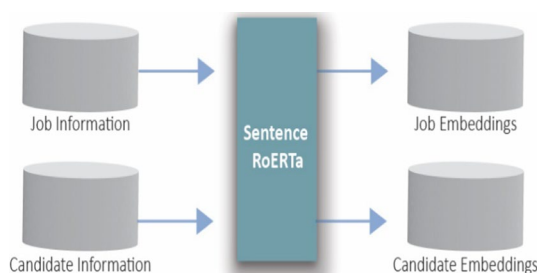


Fig. 10 High-level design—embeddings generation

TF-IDF has been chosen as a vectorise approach, as, in this case, the count vectorisation approach will perform better due to the nature of the required filters. Then the third task is to feed these vectors to the birch algorithm for the clustering task, then send the clustered groups to the DL. The goal of this layer is to cluster groups job specialisations with group ID, rather than dealing with all data points, the second layer will map data inside each group ID only. The code we used is available at: https://github.com/AleisaMonirah/Initial_Screening_Layer/blob/main/Initial_Screening_Layer.ipynb.

Mapping layer

This layer deals with the job groups produced from the layer above. Both the recruiter and the job seeker sides are given a cluster ID. We made some improvements in this layer after we start implementing it. This layer, mentioned in our last paper (Aleisa et al., 2021), was supposed to implement Word2Vec to convert the words in the job description and the job seeker’s qualification to numeric values, which are vectors; then store these values in the DL and retrieve then when needed. Then to check the similarity of the words in the job seeker’s qualification to the dictionary of words built for each job ID. This calculation was for both sides. It was not a redundancy task. Any

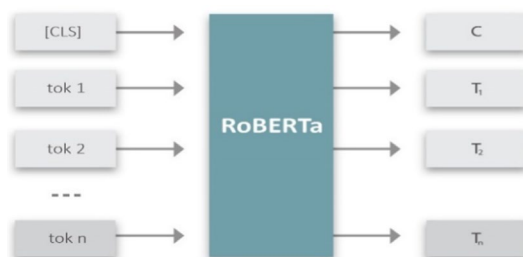


Fig. 11 RoBERTa input schema adapted by the authors from Narayanaswamy (2021)

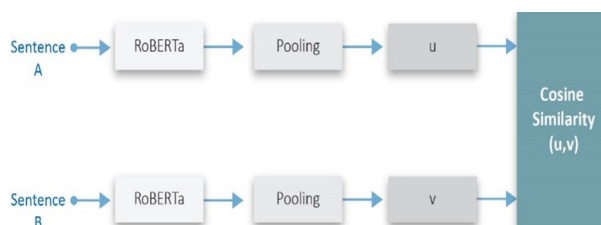


Fig. 12 Sentence transformers RoBERTa adapted by the authors from Reimers and Gurevych, (2019)

outliers were detected and removed. Then, the vectors of each job ID and job seeker ID were stored in a DL as a data frame. The improvement is that we used sentence transformers with RoBERTa as a base model (Reimers & Gurevych, 2020) instead of Word2Vec due to its robustness. For sentence transformers with RoBERTa as a base model, a high-level design embeddings generation is shown in Fig. 10. Regarding FAISS to index, FAISS is a similarity search algorithm developed by Facebook (Johnson et al., 2019). FAISS uses a data structure called index; and can scale up to billions of vectors that possibly do not even fit in Random Access Memory (RAM).

Figure 11 shows how the input is fed to sentence transformers with RoBERTa as a base model. The input text is tokenised using the Roberta specific tokeniser. RoBERTa uses a byte-level BPE tokeniser. The tokenised text is concatenated with unique tokens such as [CLS] (unique classification token) and [SEP] (start of a sentence token) at the beginning and end of sentences. It is done to ensure that the model understands the beginning and the end of the text. Positional embedding is also appended to input tokens to keep track of the distance between different tokens. The output of the RoBERTa model consists of vectors corresponding to each input token, including [CLS] and [SEP]. Each output vector has a dimension of 768. These vectors can be used for downstream tasks such as text classification, similarity, question-answering and clustering.

Since the length of input tokens varies in each document, the number of output vectors also varies consequently. In order to overcome this, RoBERTa embeddings for whole input can be averaged, or an output vector corresponding to a [CLS] token can also be chosen for further analysis. However, one study shows that sentence embeddings obtained from averaging embeddings or [CLS] tokens are unfeasible for use with cosine similarity measures due to poor performance (Reimers & Gurevych,

2019). This research aims to find similar job–candidate pairs, so embeddings needed to be optimised specifically for textual similarity tasks. In an attempt to achieve that, sentence transformers were used in this research. A sentence transformer consists of a transformer-based Siamese network. Siamese networks are neural networks with several sub-networks that have the same architecture and weight. In this research, this sub-network refers to the RoBERTa model; see Fig. 12.

Input sentences are fed to the RoBERTa model, sharing the same weights. The number of output vectors from models is different, depending on the length of input text. Therefore, a pooling strategy is used to obtain sentence embeddings. These embeddings are used to optimise the regression function, which is cosine similarity. Studies show that sentence transformers have SoA performance on a semantic textual similarity (STS), STS-benchmark dataset (Reimers & Gurevych, 2019). Jobs and candidates data are fed into sentence transformers with the RoBERTa model. The output consists of 768 dimensions embeddings corresponding to each job and each candidate.

Similar job–candidate retrieval architecture is shown in Fig. 13. The traditional cosine similarity can be very slow, as the order of time complexity is $O(m*n)$, where m and n are the number of documents to be compared for cosine similarity. In order to tackle this problem, FAISS is used for a similarity search. It also provides support for GPU supported document retrieval, and solves the scalability and latency issues as explained above. FAISS uses Euclidean distance to perform each operation on given embeddings. The output from FAISS is the top N similar documents (jobs or candidates). Apart from similar documents, it also provides distance scores that are inversely proportional to similarity: the higher the Euclidean distance, the less the similarity and vice versa.

In the AIRM study, a FAISS index was created as a second task in the mapping layer, and all the embeddings as generated vector space models were added to this index. Once the index was fixed, similar N jobs could be found for a given candidate along with a similarity measure score. Similarly, top N candidates could be identified for a given job posting. The goal of this layer is to retrieve the optimal top ten match candidates for the recruiter and the top ten available job matches for the candidate. The code we used is available at: https://github.com/AleisaMonirah/Mapping_layer/blob/main/Mapping_Layer_Code.ipynb.

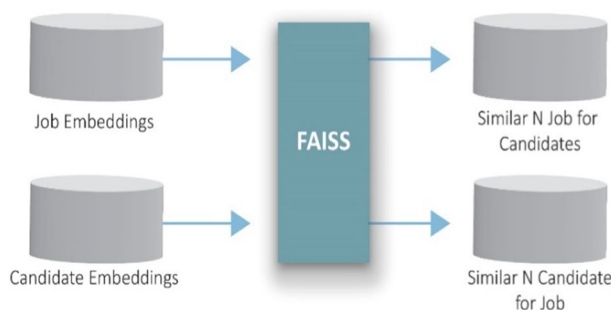


Fig. 13 High-level design—similar job–candidate retrieval

Preferences layer

This layer adds the user's preferences using the pre-trained cross-encoders model to re-rank the result list, while considering the weight of the more important words for users from both sides. Pre-trained cross-encoders take a text pair as input and provide a score ranging from 0 to 1. They do not compute embeddings for particular texts or work for individual sentences (SBERTdocuments, 2021). Cross-encoders are different from bi-encoders in the way that both sentences are passed through the model simultaneously, whereas, in bi-encoder architecture, sentences are passed separately from models sharing the same weights, as shown in Fig. 10. Cross-encoders are suitable for tasks when similarity needs to be computed between pre-defined datasets. However, cross-encoders do not provide embeddings and hence are not suitable for tasks like clustering. Bi-encoders are more suitable for tasks such as information retrieval semantic searches, as embeddings can be computed with the help of encoders which are optimised for regression loss.

The user (recruiter or job seeker) will enter keywords as preferences for the degree, the experience and general preferences. The model will give scores for the re-ranking, and these scores are combined and normalised to give the final output. For example, if a recruiter is looking for someone who has worked in an academic field and has published research papers, a candidate with a publications record has a higher score. If the recruiter wants this candidate with a specific number of years of experience, the recruiter will enter the year's segment to give more weight to a candidate with such experiences. Then cross-encoders will be applied, and the result will be stored back in the DL. The goal of this layer is to re-rank the top ten matches after adding preferences scores to the calculations. Adding the user preference is the last part of the ranking process. As a job seeker or recruiter, the user will enter some words that will re-rank the mapping layer's shortlist for them. In this layer, semantic search as embeddings can be computed with the help of encoders which are optimised for regression loss. The user (recruiter or job seeker) will enter keywords as preferences as preferences for type and level of degree, experience, or other characteristics. The model will give scores for the re-ranking, and these scores are combined and normalised to give the final output. Pre-trained cross-encoders require the input of a text pair and output a score of 0..1. This layer conducts the final procedure for the shortlist produced in the mapping layer. To show the user the best suitable match for them, after adding their preferences. The more preference words, the more complicated it gets. The code we used is available at: <https://github.com/AleisaMonirah/Preference-Layer/blob/main/Preference%20Layer.ipynb>.

Results

The AIRM was evaluated as an experiment on ten jobs sampled from the job database. Tasks show a superior quality while being more parallelisable and requiring significantly less time. The top ten candidates for each job posting were retrieved using the algorithms proposed in Sect. 7. In order to ensure the quality of similar candidate retrieval, the top 10 similar documents were retrieved. We evaluated the AIRM by using two metrics: first, the accuracy and then the time. Accuracy for each category was calculated with the help of the equation below:

Table 3 HR judgement on AIRM results job-to-candidate

Job	Candidate	HR1	HR2	HR3	Final
Job 1	Candidate 1	Good	Good	Good	3
	Candidate 2	Good	Good	Good	3
	Candidate 3	Good	Good	Good	3
	Candidate 4	Good	Good	Good	3
	Candidate 5	Good	Good	Good	3
	Candidate 6	Good	Good	Good	3
	Candidate 7	Good	Good	Good	3
	Candidate 8	Good	Good	Good	3
	Candidate 9	Good	Bad	Good	2
	Candidate 10	Bad	Good	Bad	1
Job 10	Candidate 1	Good	Good	Good	3
	Candidate 2	Good	Good	Good	3
	Candidate 3	Good	Good	Good	3
	Candidate 4	Good	Good	Good	3
	Candidate 5	Good	Good	Good	3
	Candidate 6	Good	bad	Bad	1
	Candidate 7	Bad	Good	Bad	1
	Candidate 8	Bad	Bad	Good	1
	Candidate 9	Bad	Bad	Bad	0
	Candidate 10	Bad	Bad	Bad	0

The complete result is available at: https://github.com/monirahaleisa/AIRM-Results-Job-to-Candidate/blob/main/Mapping_results.pdf

Table 4 AIRM results job-to-candidate

	Categories	Per cent
Three agreed	3	61
Two agreed	2	5
One agreed	1	18
No one agreed	0	16
Total		100

$$\text{Accuracy} = \frac{\text{Number of correct candidates retrieved by the algorithm}}{\text{Total number of candidates retrieved by the algorithm}}$$

HR experts worked manually as a research collaboration to evaluate the results of the AIRM for the ten sampled jobs by assessing the top ten documents retrieved for the ten similar candidates chosen for each job. Human intervention was required to assess the algorithm's performance for two reasons: firstly, we use unsupervised algorithms; secondly, choosing a candidate for a job is a subjective judgement.

For the evaluation task, the predictions were labelled as 'category three' if all of the three evaluators agreed on them; 'category two' if two evaluators agreed on them; and 'category one' if only one evaluator thought it was a good match; otherwise, they have labelled them as 'category zero', as shown in Table 3.

Table 5 Time taken by HR experts to finish the task

	Time in days
HR1	6
HR2	5
HR3	7
Total	18
Average	6

From the AIRM results, 61% of the matches falls into category three, 5% of the matches falls into category two, 18% of the matches fell into category one, and 16% fell into category zero.

As demonstrated in Table 4, the AIRM system gives an overall 84% accuracy of matching, with at least one expert agreeing with the system's selection. See the formula above for calculating accuracy. Therefore, we consider the AIRM system to be suitable for the automatic pre-selection of candidates matching the job description, ready for further refinement by human experts.

As shown in Table 5, the average time it took HR experts to analyse the matching of 100 candidates to 10 jobs was 6 days, or roughly 5 h each day; each HR expert calculated how long they worked on each candidate's CV to the minute. Additional file 1: Appendix Example of HR consideration when matching CV to jobs.

In a test that contained 2012 records of jobs and 16,255 candidate records, the AIRM exceeded human performance in terms of time. It finished the work in 2.4 min, whereas humans spent approximately 6 days on average for only 10 jobs. This means that the AIRM is beneficial for pre-selecting a block of potential candidates for a particular job. We are optimistic about the future of the AIRM and intend to use it for a variety of purposes. We intend to expand the AIRM to solve challenges, including combining two sides of the text and dealing with massive inputs and outputs. One of our future is to implement the AIRM to work with the Arabic language and compare the results with the English language version.

The model is supposed to apply ML and AI techniques to assist in matching job seekers to vacant positions. Several things became clear to us as we tried to understand the model's results and behaviour:

- The model does not credit the candidate for having a double major. In contrast, the recruiters consider having different majors as an advantage. They believe that such candidates have expanded their prospects and have more knowledge.
- The model does not give candidates with master's degrees more credit than candidates with bachelor's degrees. Hence the importance of having a preference layer, as a preference layer is critical in such cases.
- In cases where the candidate had worked in different jobs, the model fails to recognise that this individual's skills has evolved, and they should be categorised as experienced rather than as a beginner, and can therefore be a candidate for intermediate or senior jobs.
- The model cannot capture abbreviations easily, such as the acronym for Master of Business Administration (MBA).

- In cases where the job has no matches at all in the DL, the model will still bring forward the top ten candidates with the least cosine similarity distance. We cannot specify a specific distance as a benchmark to stop fetching data at some point. In this case, the model will seem to be bringing in illogical candidates. This point can be addressed by future research that could explore when the model could stop requesting data by attempting to locate or compute a stop benchmark distance.
- Some notions, such as transferable skills, are not recognised by the model. If a candidate has managerial expertise in one sector; then it can be assumed that these talents may be utilised in another sector. The model fails to recognise that a specific talent may be generalised and utilised in many sectors.

In addition to the aforementioned points, these results suggest more scope for future research. The researcher can see a promising future in implementing the AIRM to work with Arabic and comparing the results with the English version. Moreover, building the CV with blockchain technology to increase trust in the CV.

Justification of the implemented solution

The following components were chosen over others due to the given reasons:

DL has many advantages over old storage systems, from both business and technological perspectives. Businesses benefit from DLs due to their all-around data availability or democratisation and because they effortlessly fetch good quality data. Technological benefits from DLs are real-time decision analysis, SQL and other languages support, scalability and versatility. The following are the most commonly considered advantages of DLs (Chinnakali, 2016; DATABRICKS, 2020; Warren, 2019):

- Scalability, which can handle a growing amount of data.
- A DL can store all types of data: logs, extensible markup language (XML), multimedia, sensor data, binary, social data, chats and people data.
- A DL can leverage data to high-speed data, integrating with the historical data to have its most exclusive insights.
- DLs implant the schema to leverage both structured and unstructured data.
- In DLs, modelling is required only during data consumption and not during ingestion.
- DLs leverage Hadoop's simplicity to store data based on schema-less write and schema-based read modes, which are highly applicable during data consumption.
- DL excels at utilising the availability of large quantities of coherent data along with deep learning algorithms to recognise items of interest that will power real-time decision analytics.

BIRCH (balanced iterative reducing and clustering using hierarchies) is suitable for AIRM for four reasons:

- BIRCH clustering does not try to make clusters of the same size as in k-mean clustering. The group size will vary in the Saudi jobs data set, and we do not aim to have the same group size. Instead, we need to explore the actual group size in each job group (Zhang et al., 1997).
- BIRCH clustering does not require the number of clusters as an input parameter. In the Saudi jobs data set, the data are significant, and it will change all the time. We also cannot decide the number of clusters at the beginning of the algorithm. So hierarchical clustering helps to take away the problem of having to pre-define the number of clusters (Lorbeer et al., 2018; Pathak, 2018).
- BIRCH clustering can handle any distance metric virtually (SHARMA, 2019).
- BIRCH has been proposed for minimising the running time of clustering. BIRCH incrementally clusters enormous datasets whose sizes are much greater than the amount of available memory. The clustering process is performed by constructing a height-balanced tree (Zhang et al., 1997).

The sentence transformer method captures semantic similarity better than other methods because it is trained in the Siamese network fashion. The output loss is cosine similarity which is optimised. In this way, the model learns to capture similarity better than other models and perform SoA on the semantic textual similarity (STS) benchmark dataset. The Robustly Optimised Bert Pretraining Approach (RoBERTa) is part of Facebook's continued effort to push the boundaries of self-supervised systems that can be built with less reliance on time- and resource-intensive data labelling.

FAISS (Facebook AI Similarity Search) In order to compute the cosine similarity between embeddings, traditional cosine similarity leads to enormous computation resources and has a time complexity of $O(n*m)$. This complexity is reduced by using index-based data structures, which FAISS inherently supports. In this way, similar documents can be retrieved very fast. FAISS is also scalable to up to billions of vectors and can search out similarity even for embeddings that do not fit in random access memory (RAM).

Pre-trained cross-encoders are suitable for a pre-defined set of sentence pairs, as in this research, and they only need to be scored; for example, when 100 sentence pairs need to get similarity scores for these 100 pairs. Cross-encoders are slower than bi-encoders; however, they are suitable for computing the sentence embeddings needed in this research, which is not a huge number.

Conclusion

This section will be concluding the paper from three dimensions: first, the theoretical implications, then the managerial implications, and finally, future research.

In order to account for the subjective nature of the selection process when solely conducted by human HR experts, this study AIRM system. The AIRM was evaluated based on two metrics: accuracy and time. The results showed that the AIRM achieved an overall matching accuracy of 84%, with at least one expert concurring with the system's output. Additionally, the AIRM completed the task in 2.4 min, while human experts took an average of more than 6 days. These findings indicate that the AIRM

outperforms human experts in terms of task execution, highlighting its potential value in pre-selecting applicants and positions. Importantly, the AIRM's applicability extends beyond government services and can be beneficial for any commercial business utilising Big Data.

Theoretical implications

This paper continues the work of our recent paper (Aleisa et al., 2021), where we proposed an AIRM architecture in order to assist the labour market. In the current paper, we implemented the MVP of the proposed AIRM architecture and found impressive results, where we exploit ML and AI power. The AIRM architecture consists of a data repository, the DL and three processing layers that use ML and AI. The DL lays in the heart of AIRM architecture, the three layers of models that are stacked on top of each other feed data into and out of the DL.

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 by using BIRCH. This ensures search latency is reduced. Jobs and candidates are clustered together. For example, 20 clusters in total. If the candidate belongs to the fifth cluster, then a job will also be searched for him or her in the fifth cluster. This ensures latency while searching for similar jobs.

The second layer is called a mapping layer, where the sentence transformers with RoBERTa as a base model is used for transfer learning. This transformer-based model is used to generate embeddings for job descriptions, as well as candidate profiles. It is optimised to find similar documents. Once embeddings are generated, then these embeddings are indexed in FAISS. FAISS is used to perform an approximate nearest neighbour search and has very low latency. It is easily scalable as well. Using FAISS, we can find an 'n' number of similar jobs for a candidate and vice versa. FAISS overcomes the computational expense of traditional cosine similarity.

The third layer is the preferences layer, which will add the preferences as the weight of a word that is more important for both sides. Here we used the pre-trained cross-encoders model to re-rank the result list, while considering the more important words for users from both sides. Pre-trained cross-encoders take a text pair as input and provide a score ranging from 0 to 1. Then the result will be stored back in the DL.

In order to evaluate the algorithm's performance, it was necessary to use human input. As part of the evaluation task, three human experts in HR and recruiting evaluated the results for ten selected jobs, with ten matching candidates for each job. We considered appropriate an AIRM selection agreement by at least one human expert to accommodate the subjective nature of the selection process when performed entirely by human HR and recruitment experts. We found that the AIRM system gave an overall 84% accuracy of matching, with at least one expert agreeing with the system's selection. This result may change due to the subjective nature of the task, and another set of HR experts may give different opinions. Additionally, the model's behaviour needs to be tested again if more data are collected. The AIRM finished the work in 2.4 min, whereas humans spent

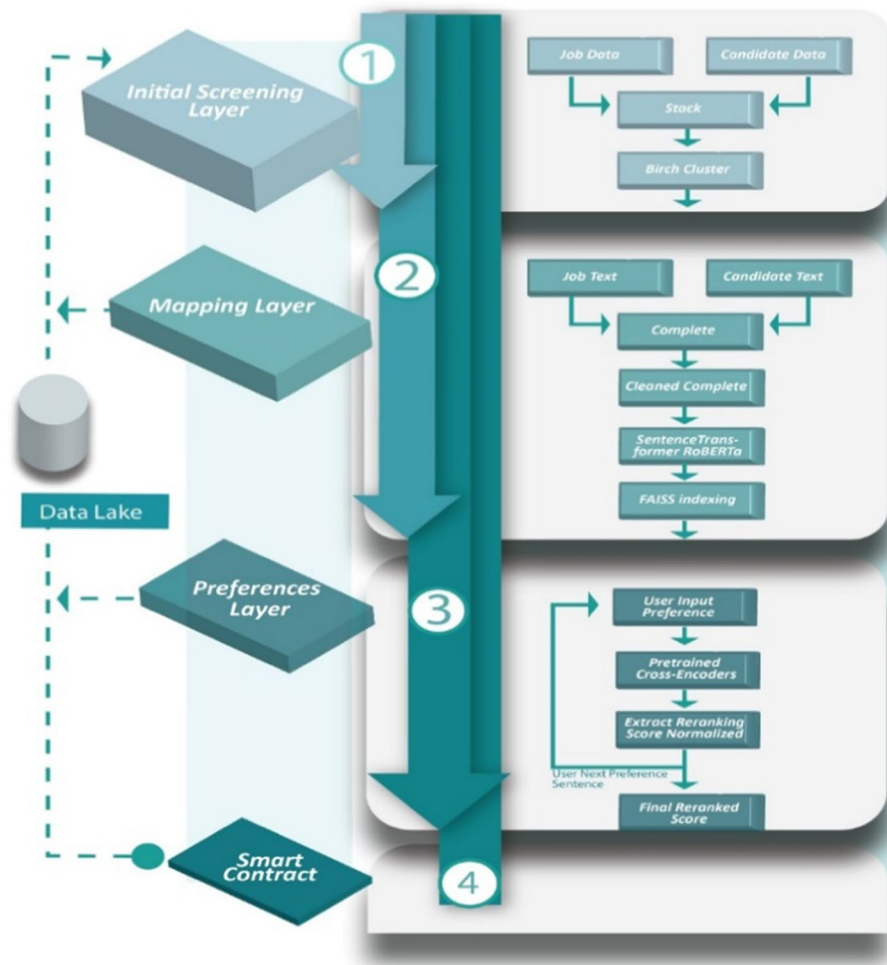


Fig. 14 Smart contract added to AIRM

more than six days on average, which is beneficial in pre-selecting a block of candidates and jobs. It is the first time a single model has reached a new state-of-the-art in overall performance.

Managerial implications

Due to the nature of the research and, as with any data-driven project, the researcher needed domain experts to collaborate in analysing the model's results. To conduct this analysis, the researcher could not stand alone, as it was necessary to use expert human analysis in order to evaluate the AIRM's performance. We considered that an AIRM selection agreed by at least one expert accommodated the subjective nature of the selection process when performed entirely by human HR and recruitment experts. In other words, agreement by one expert was considered to be sufficient to indicate an appropriate match between the job specification and the candidate. We found that the AIRM system gave an overall accuracy of 84%, whereby at least one expert agreed with the system's selection.

The AIRM finished the work in 2.4 min, whereas humans spent an average of 6 days, which demonstrated the time-saving benefits of using AIRM in pre-selecting a block of candidates suitable for a specific job. Therefore, the researchers consider the AIRM system to be suitable for the automatic pre-selection of candidates matching the job description, to allow human experts to concentrate on more detailed and nuanced consideration of an already pre-selected subset of job seekers, thus improving the efficiency of human involvement.

Future research

This research contributes to the field of mapping human-written texts. Furthermore, it paves the way for academics to use the AIRM and blockchain technologies to improve trust in AI by the use of blockchain smart contracts. This can be achieved by utilising blockchain technology to create CVs and saving them in DL to match job openings via the AIRM. The AIRM can be expanded to support languages other than English. The sentence transformer model of the AIRM could be pre-trained in other languages, while the rest of the AIRM architecture would remain the same. The AIRM can be used to support commercial services.

This research focuses on government services, but the AIRM architecture can be used in any enterprise that utilises big data. Further work for this research project is to find different types of data sets in order to generalise the AIRM for other government uses. Moreover, we plan to implement the AIRM using the Arabic language and evaluate its performance. More interesting further work is to add a fourth layer to the AIRM to embed AI into blockchain smart contracts, to integrate AI and blockchain capabilities; see Fig. 14.

This will enable a more in-depth examination of the efficacy of the contract's terms and the procedures it governs. Consequently, human analysis, intervention and verification are considerably minimised. The cutting-edge, dynamic combination of AI and blockchain significantly simplifies the negotiation, the execution process and builds more trust in the AIRM. Because AI performs better when data are collected through a reliable, secure, trustworthy and credible data repository or platform, trust in the AIRM would develop. Blockchain is a distributed ledger in which data may be cryptographically signed, authenticated and agreed upon by all mining nodes. The combination of AI and blockchain-enabled smart contracts would create business solutions that could build on existing enterprise systems while also adjusting them in favour of next-generation alternatives.

Abbreviations

AIRM	Artificial Intelligence Recruiting Model
AI	Artificial Intelligence
API	Application Programming Interface
ASC	Arab Standard Classification
BiLSTM	Bidirectional Long-Short Term Memory network
BERT	Bidirectional Encoder Representations from Transformers
BI	Business Intelligence
BIRCH	Balanced Iterative Reducing and Clustering using Hierarchies
BoW	Bag of Words
BSS	Business Support Systems
CRISP-DM	Cross Industry Standard Process for Data Mining
CRM	Customer Relationship Management
CPU	Central Processing Units
CNN	Convolutional Neural Networks
CV	Curriculum Vita
CzRM	Citizen Relationship Management

DM	Data Mining
DL	Data Lake
DW	Data Warehouse
EDA	Data Exploration Analysis
ELT	Extract, Load, Transform process
ETL	Extract, Transform, Load process
FAISS	Facebook AI Similarity Search
GDP	Gross Domestic Product
GPU	Graphic Processing Units
GOSI	General Organization for Social Insurance
GRU	Gated Recurrent Unit
GSTAT	General Authority for Statistics
HR	Human Resource
HADAF	Human Resources Development Fund
ILO	International Labour Organization
ISCO	International Standard Classification of Occupations
KDD	Knowledge Discovery in Databases
LSTM	Long-Short Term Memory Networks
MCS	Ministry of Civil Service
ML	Machine Learning
MLSD	Ministry of Labour and Social Development
MT	Machine Translation
MVP	Minimum Viable Product
NDIC	National Digital Information Centre
NIC	National Information Centre
NLI	Natural Language Inference
NLO	National Labour Observatory
NLP	Natural Language Processing
NSA	National Security Agency
OLTP	Online Transaction Processing
OSS	Operations Support Systems
RNN	Recurrent Neural Networks
RoB-ERTa	Robustly Optimised BERT Pretraining Approach
SBERT	Sentence-BERT
SEMMA	Sample, Explore, Modify, Model, Assess.
SDAIA	Saudi Data and Artificial Intelligence Authority
SoA	State of the Art
STS	Semantic Textual Similarity
SQL	<i>Structured Query Language</i>
TF-IDF	Term Frequency–Inverse Document Frequency

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13731-023-00324-w>.

Additional file 1: Example of HR consideration when matching CV to jobs.

Acknowledgements

This paper and its research would not have been possible without the three HR experts who devoted their time to giving the research their opinions: Mr. Yaser Arafath, Mrs. Rana Alwetaid and Miss. Hessah Alfozan.

Author contributions

MA—as main author, is the major contributor of all scientific aspects of this paper based on her PhD research topic; NB—the main supervisor of this research project, contributed to the overall research planning, structure and style of the paper and main discussion/conclusion, and scientific argument; MW—the second supervisor of this research project, contributed to the structure and style of the paper and the main discussion/conclusion, and scientific argument.

Funding

Self-funded.

Availability of data and materials

All data are available upon request.

Declarations

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 11 October 2021 Accepted: 17 August 2023

Published online: 14 September 2023

References

- Affairs_of_V2030. (2016). *V2030 5th Anniversary-vision 2030*. EconomicAffairs. <https://www.vision2030.gov.sa/ar/media-center/news/v2030-5th-anniversary/>
- Aggarwal, C. C., & Reddy, C. K. (2014). *Data clustering: algorithms and applications* (pp. 234–297). Boca Raton, Fla.: CRC Press ProQuest Ebook.
- Albaker, A., & Alabdani, A. (2018). The challenges of the labour market in the Kingdom of Saudi Arabia. *Saudi Monetary Agency, in Arabic*.
- Aleisa, M. A., Beloff, N., & White, M. (2021). AIRM: A new AI recruiting model for the Saudi Arabia Labour Market. *Intelligent Systems Conference (IntelliSys)*, 2021(296), 105–124. https://doi.org/10.1007/978-3-030-82199-9_8
- Almaoasi, M. (2017). Crown Prince puts it on the challenge and is expected to reap the rewards. *SABQ*. <https://sabq.org/vTgSKr>
- Alsultanny, Y. A. (2013). Labour market forecasting by using data mining. *Procedia Computer Science*, 18, 1700–1709. <https://doi.org/10.1016/j.procs.2013.05.338>
- Al-Zughaibi, S. A. (2014). *The importance of labour market reforms in restructuring the Saudi economy*. <https://www.alriyadh.com/915519>
- Apache NiFi Team. (2021). *Apache NiFi Documentation*. <https://nifi.apache.org/docs.html>
- Arya Leoforce. (2021). *Arya Quantum-AI Powered Sourcing Solution, Integrated into Your Existing ATS*. Goarya.Com. <https://goarya.com/arya-quantum/>
- Bajaj, P., Campos, D., Craswell, N., Deng, L., Gao, J., Liu, X., Majumder, R., Mcnamara, A., Mitra, B., Nguyen, T., Rosenberg, M., Song, X., Stoica, A., Tiwary, S., & Wang, T. (2018). MS MARCO: A human generated machine reading comprehension dataset. <https://www.jeopardy.com/>
- Barenkamp, M., Rebstadt, J., & Thomas, O. (2020). Applications of AI in classical software engineering. *AI Perspectives* 2020:1, 2(1), 1–15. <https://doi.org/10.1186/S42467-020-00005-4>
- Boselli, R., Cesarini, M., Mercurio, F., & Mezzanzanica, M. (2017). Using Machine Learning for Labour Market Intelligence. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10536 LNAI, 330–342. https://doi.org/10.1007/978-3-319-71273-4_27
- Chala, S. A., Ansari, F., & Fathi, M. (2016). A framework for enriching job vacancies and job descriptions through bidirectional matching. *WEBIST 2016-Proceedings of the 12th International Conference on Web Information Systems and Technologies*, 2(Webist), 219–226. <https://doi.org/10.5220/0005806502190226>
- Chala, S. A. (2018). Bidirectional job matching through unsupervised feature learning. In *Universität Siegen* (pp. 1–161).
- Chinnakali, K. (2016). *9 Key Benefits of Data Lake-Data Science Central*. <https://www.datasciencecentral.com/profiles/blogs/9-key-benefits-of-data-lake>
- Colombo, E., Mercurio, F., & Mezzanzanica, M. (2018). Applying machine learning tools on web vacancies for labour market and skill analysis.
- Copeland, B. J. (2017). *Artificial intelligence|Definition, Examples, and Applications|Britannica*. Britannica.Com. <https://www.britannica.com/technology/artificial-intelligence>
- DATABRICKS. (2020). *Introduction to Data Lakes-Databricks*. Databricks.Com. <https://databricks.com/discover/data-lakes/introduction>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv*. <https://github.com/tensorflow/tensor2tensor>
- el Naqa, I., Ruijiang, L., & Martin, J. M. (2015). Machine learning in radiation oncology. In *Machine Learning in Radiation Oncology*. <https://doi.org/10.1007/978-3-319-18305-3>
- Fang, H. (2015). Managing data lakes in big data era: What's a data lake and why has it become popular in data management ecosystem. *2015 IEEE International Conference on Cyber Technology in Automation, Control and Intelligent Systems, IEEE-CYBER 2015*, 820–824. <https://doi.org/10.1109/CYBER.2015.7288049>
- Fayyad, U. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3). https://doi.org/10.1007/978-3-319-18032-8_50
- Frades, I., & Matthiesen, R. (2010). Overview on techniques in cluster analysis bioinformatics methods in clinical research. 593, 81–107. <https://doi.org/10.1007/978-1-60327-194-3>
- François, D. (2008). Methodology and standards for data analysis with machine learning tools. *ESANN 2008 Proceedings, 16th European Symposium on Artificial Neural Networks - Advances in Computational Intelligence and Learning, January 2008*, 239–246.
- General Authority for Statistics. (2018). *Labour Force|General Authority for Statistics*. <https://www.stats.gov.sa/en/814>
- General Authority for Statistics. (2021). Labour market statistics for the first quarter of 2021. *General Authority for Statistics*, 2021.
- Gill, K., Scott, E., & Ward, L. (2015). Understanding labour market information. *Cambridge Training and Development*.
- Hall, R. P., & Kibler, D. F. (1985). Differing methodological perspectives in artificial intelligence research. *AI Magazine*, 6(3), 166–178.
- Harvard Kennedy School. (2021). *The Labour Market in Saudi Arabia*. https://epod.cid.harvard.edu/sites/default/files/2019-08/EPD_Report_Digital.pdf
- Homenda, W., & Pedrycz, W. (2018). CLUSTERING. In *Pattern Recognition* (pp. 247–273). Wiley. <https://doi.org/10.1002/9781119302872.ch8>
- Ikudo, A., Lane, J., Staudt, J., & Weinberg, B. A. (2018). Occupational classifications: A machine learning approach. *SSRN Electronic Journal*, 11738. <https://doi.org/10.2139/ssrn.3229805>

- Janrao, P., & Palivela, H. (2015). Management zone delineation in Precision agriculture using data mining: A review. *ICIIECS 2015-2015 IEEE International Conference on Innovations in Information, Embedded and Communication Systems, September*. <https://doi.org/10.1109/ICIIECS.2015.7193256>
- Janzz. (2021). *JANZZ technology: Leader in semantic skills and job matching*. <https://janzz.technology/>
- Johnson, J., Douze, M., & Jegou, H. (2019). Billion-Scale Similarity Search with GPUs. *IEEE Transactions on Big Data*, 7(3), 535–547. <https://doi.org/10.1109/TBDATA.2019.2921572>
- Kenton, W. (2020). *Okun's Law Definition*. Investopedia. <https://www.investopedia.com/terms/o/okunslaw.asp>
- Khine, P. P., & Wang, Z. S. (2018). Data lake: A new ideology in big data era. *ITM Web of Conferences*, 17(December), 03025. <https://doi.org/10.1051/itmconf/20181703025>
- Larsson, A., & Teigland, R. (2020). *The Digital Transformation of Labour: Automation, the Gig Economy and W*. <https://www.routledge.com/The-Digital-Transformation-of-Labor-Automation-the-Gig-Economy-and-Welfare/Larsson-Teigland/p/book/9780367330705>
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., & Allen, P. G. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv*, 1907.11692. <https://github.com/pytorch/fairseq>
- Lorbeer, B., Kosareva, A., Deva, B., Softić, D., Ruppel, P., & Küpper, A. (2018). Variations on the clustering algorithm BIRCH. *Big Data Research*, 11, 44–53. <https://doi.org/10.1016/j.bdr.2017.09.002>
- Mareckova, J., & Pohlmeier, W. (2017). Noncognitive skills and labor market outcomes: a machine learning approach. *Beiträge. Econstor Make Your Publications Visible, G03-V2*. <http://hdl.handle.net/10419/168195>
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *ArXiv*, 1301.3781. <http://ronan.collobert.com/senna/>
- Mishra, S., & Tripathi, A. R. (2021). AI business model: An integrative business approach. *Journal of Innovation and Entrepreneurship*, 10(1), 1–21. <https://doi.org/10.1186/S13731-021-00157-5/FIGURES/13>
- National Digital Transformation. (2020). *National Digital Transformation report 2019*. <https://ndu.gov.sa/sites/default/files/2021-09/Biannual-report-en.pdf>
- Narayanaswamy, G. R. (2021). *Exploiting BERT and RoBERTa to Improve Performance for Aspect Based Sentiment Analysis* Gagan Reddy Narayanaswamy. <https://doi.org/10.21427/3w9n-we77>
- O'Brien, J. (2017). The definitive guide to the Osce. *Elsivier*, 110(9), 1689–1699.
- Pathak, M. (2018). *Hierarchical Clustering in R-DataCamp*. Datacamp. <https://www.datacamp.com/community/tutorials/hierarchical-clustering-R#what>
- Privatization Program. (2018). *Privatization Program-Saudi Vision 2030*. <https://www.vision2030.gov.sa/ar/v2030/vrps/privatization/>
- Qaiser, S., & Ali, R. (2018). Text mining: use of TF-IDF to examine the relevance of words to documents. *International Journal of Computer Applications*, 181(1), 25–29. <https://doi.org/10.5120/ijca2018917395>
- Quix, C., & Hai, R. (2018). Data Lake. *Encyclopedia of Big Data Technologies*, 1–8. https://doi.org/10.1007/978-3-319-63962-8_7-1
- Rani, Y., & Rohil, H. (2013). A study of hierarchical clustering algorithm. In *International Journal of Information and Computation Technology* (Vol. 3, Issue 10). <http://www.irphouse.com/ijict.htm>
- Reich, Y. (1994). Layered models of research methodologies. *Artificial Intelligence for Engineering, Design, Analysis and Manufacturing*, 8(4), 263–274. <https://doi.org/10.1017/S0890060400000949>
- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *ArXiv*.
- Reimers, N., & Gurevych, I. (2020). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *EMNLP-IJCNLP 2019-2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 3982–3992. <https://doi.org/10.18653/v1/d19-1410>
- SBERT documents. (2021). *Pretrained Cross-Encoders—Sentence-Transformers documentation*. SBERT.NET. https://www.sbert.net/docs/pretrained_cross-encoders.html
- Scarpetta, S., & Sonnet, A. (2012). Challenges facing European labour markets: Is a skill upgrade the appropriate instrument? *Intereconomics*, 47(1), 4–30. <https://doi.org/10.1007/s10272-012-0402-2>
- SDAIA. (2021). *SDAIA*. <https://sdaia.gov.sa/>
- SHARMA, P. (2019). *Hierarchical Clustering | Hierarchical Clustering Python*. Analyticsvidhya. <https://www.analyticsvidhya.com/blog/2019/05/beginners-guide-hierarchical-clustering/>
- Sundsøy, P., Bjelland, J., Reme, B. A., Jahani, E., Wetter, E., & Bengtsson, L. (2017). Towards real-time prediction of unemployment and profession. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10540 LNCS, 14–23. https://doi.org/10.1007/978-3-319-67256-4_2
- Think Big. (2018). *Kylo Documentation*.
- van der Zande, J., Eigland, K., Shahryar, S., & Teigland, R. (2020). The substitution of labour: from technological feasibility to other factors influencing the potential of job automation. In *The Digital Transformation of Labour: Automation, the Gig Economy and Welfare* (Issue January).
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 2017-Decem* (Nips), 5999–6009.
- Walch, K. (2020). *Why Agile Methodologies Miss the Mark for AI & ML Projects*. Forbes Media LLC. <https://www.forbes.com/sites/cognitiveworld/2020/01/19/why-agile-methodologies-miss-the-mark-for-ai-ml-projects/?sh=307b979e21ea>
- Warren, J. (2019). *The Business and Technological Benefits of Data Lakes*. <https://www.smartdatacollective.com/business-and-technological-benefits-of-data-lakes/>
- WAS_SPA. (2017). *National Information Center launches its digital strategy*. Saudi Press Agency. <https://www.spa.gov.sa/1582613>
- Wiggers, K. (2018). *Developers can now use Google's Cloud Talent Solution to power job searches* | VentureBeat. <https://venturebeat.com/2018/08/16/2384861/>
- Wilkinson, D. (2002). *The Researcher's Toolkit: The Complete Guide to Practitioner Research: Discovery Service for Saudia Digital Library*. In London: Routledge. Print/E-mail/Save 60 Pages.
- Wowczko, I. A. (2015). Skills and vacancy analysis with data mining techniques. *Informatics*, 2(4), 31–49. <https://doi.org/10.3390/informatics2040031>

- Zhang, X., Zhou, Y., Ma, Y., Chen, B. C., Zhang, L., & Agarwal, D. (2016). GLMix: Generalized linear mixed models for large-scale response prediction. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13–17-Aug, 363–372. <https://doi.org/10.1145/2939672.2939684>
- Zhang, T., Ramakrishnan, R., & Livny, M. (1997). BIRCH: A new data clustering algorithm and its applications. *Data Mining and Knowledge Discovery*, 1(2), 141–182. <https://doi.org/10.1023/A:1009783824328>

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- ▶ Convenient online submission
- ▶ Rigorous peer review
- ▶ Open access: articles freely available online
- ▶ High visibility within the field
- ▶ Retaining the copyright to your article

Submit your next manuscript at ▶ [springeropen.com](https://www.springeropen.com)
