Chapter 3: Research Methodology

3.1 Introduction

This chapter gives a step-by-step account of the research methodology employed to accomplish the objectives of this research, i.e., analyzing the development of skills needed in the Saudi employment market on the basis of machine learning methods. The research methodology is aimed at providing a systematic and structured methodology towards collecting, processing, analyzing, and interpreting employment market data collected through online employment portals.

Research methodology includes several steps such as data collection, preprocessing, exploratory data analysis, skills extraction via natural language processing (NLP), categorization of skills into appropriate classes, trend analysis based on time and industry, comparison of curricula at educational levels, model evaluation, and lastly conclusion and recommendation from findings.

The research design is quantitative analytical with big data analytics and AI software employed to make meaningful conclusions from unstructured text data in job advertisements.

3.2 Research Framework

The research applies a systematic framework that combines technological and analytic components. The framework facilitates correct identification and classification of skills demanded by employers across various sectors of the Saudi labor market. It also facilitates the tracking of trends over time, which informs education and workforce planning decision-making.

The framework includes the following key components:

Data Acquisition : Collecting job advertisement data from online sources like LinkedIn, Bayt.com, and GulfTalent.

Data Preprocessing : Cleaning and pre-processing the raw data to get it into analysis-ready format.

Natural Language Processing (NLP) : Extraction of major skills and abilities from unstructured job posts.

Skill Classification : Assignment of extracted skills into predefined categories like technical skills, soft skills, and business skills.

Temporal and Sectoral Trend Analysis : Determination of how skill demands change over time and vary with respect to sectors.

Comparison with Academic Curricula : Determination of the alignment between required skills and skills being supplied in Saudi universities.

Model Evaluation and Validation : Verification of the accuracy and validity of the models adopted for extracting and categorizing skills.

Strategic Recommendations : Allocation of evidence-based policy recommendations for closing the skills gap

3.3 Data Collection

3.3.1 Data Sources

In order to provide a wide coverage of recent labor market needs, job advertisements were gathered from three leading online job websites:

LinkedIn

Bayt.com

GulfTalent

They are used mainly by employers and employees in Saudi Arabia and offer access to millions of real-time job listings for different sectors of the economy including information technology, health care, engineering, finance, and public administration.

3.3.2 Keywords and Search Terms

In order to make the results effective and comprehensive, job searches were made using each industry's given key words. For instance:

Information Technology : "Data Analyst", "Cybersecurity Specialist", "Cloud Engineer"

Healthcare : "Clinical Informatics", "Medical Technologist", "Healthcare IT"

Finance : "Financial Analyst", "Risk Management", "Fintech Developer"

Job postings were restricted to those of the Kingdom of Saudi Arabia and within a certain time period to allow temporal trend analysis.

Platform	Total Ads	IT Sector	Healthcare	Finance
LinkedIn	500	200	100	120
Bayt.com	600	250	120	150
GulfTalent	400	180	90	110

Table 3.1 : A sample table showing the number of job ads collected per platform and per sector

3.4 Data Preprocessing and Exploratory Data Analysis

The gathered job descriptions were preprocessed a number of times before the applications to machine learning models were ready.

3.4.1 Text Cleaning

Unwanted characters, HTML tags, and special characters were removed from text. Duplicates or repetitive job postings were removed as well.

3.4.2 Tokenization

Text was divided into words or phrases (tokens) so that it became simpler to analyze.

3.4.3 Stop Words Removal

Arabic and English stop words (e.g., "and", "the") were eliminated to eliminate noise from the corpus.

3.4.4 Stemming and Lemmatization

Words were stemmed to their root form to normalize the lexicon. e.g., "running" → "run".

3.4.5 Part-of-Speech Tagging

Every word was tagged based on syntactic role to enhance semantic comprehension.

3.4.6 Named Entity Recognition (NER).

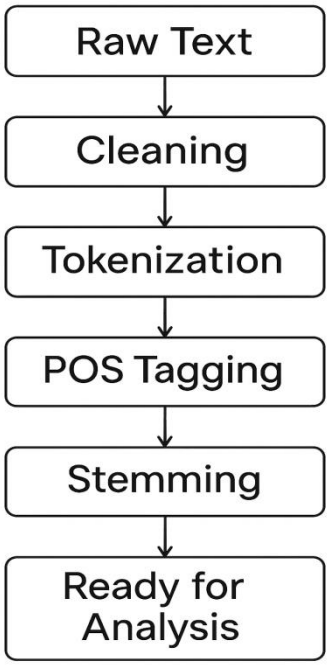

Figure 3.1 : A diagram showing the NLP pipeline

3.5 Skill Extraction and Classification

Post-preprocessing, the abilities were learned from the preprocessed job advertisements using NLP methods. The provided abilities were tagged utilizing machine learning models including Support Vector Machines (SVM), Naïve Bayes, and BERT-based models into the following categories:

Technical Skills : i.e., Python, Java, Cybersecurity, Cloud Computing

Soft Skills : i.e., Communication, Teamwork, Leadership

Business Skills : i.e., Budgeting, Project Management, Risk Assessment

Skill	Type	Required by Employers	Taught in Universities
Python Programming	Technical	□	□
Communication	Soft	□	! (Self-reported)
Project Management	Business	□	□

Table 3.2: Example of classified skills:

3.6 Time and Sector Trend Analysis

We analyzed the trend of skill demand over time and by sectors using clustering and association rule mining methods.

3.6.1 Temporal Analysis

We observed the variation in skill demand over 2020 and 2024. For example, demand for AI-proficient candidates grew strongly beyond 2022, while legacy programming languages such as Java declined.

3.6.2 Sector-Wise Analysis

Variation in skill preferences was observed in various sectors.

IT Sector : Overwhelming demand for cloud computing and cybersecurity talent.

Healthcare Sector : Focus on clinical informatics and digital health.

Finance Sector : Fintech and data analytics competencies in demand.

3.7 Model Comparison and Evaluation

Some machine learning algorithms were trained and compared for their skill to extract and classify skills from job descriptions. The metrics used were as follows:

Accuracy

Precision

Recall

F1 Score

3.7.1 Results

Model	Accuracy	F1 Score
SVM	89%	0.88
Naïve Bayes	84%	0.83
k-Nearest Neighbor	82%	0.81
BERT (Fine-tuned)	91%	0.9

Table 3.3 : comparing model performance.

3.8 Tools and Technologies Utilized

Different tools and libraries were utilized during the course of research:

Web Scraping : Octoparse, BeautifulSoup

Natural Language Processing : spaCy, NLTK, Transformers (Hugging Face)

Machine Learning Models : Scikit-learn, TensorFlow, PyTorch

Data Visualization : Matplotlib, Seaborn, Tableau

Programming Languages : Python, SQL

3.9 Limitations and Challenges

Even with the strong methodology, some challenges and limitations were faced while conducting the research:

Data Availability : There was no availability of all job postings because some restrictions were imposed on some websites.

Language Variability : Informal Arabic expressions and dialects affected accuracy in skill extraction.

Historical Data : Partial historical job postings constrained long-term trend analysis.

Model Accuracy : The novel or unusual skills were not captured very well owing to the absence of training data.

Manual Verification : Manual verification was needed for some classifications to undo misclassifications.

3.10 Conclusion

This chapter provided a comprehensive outline of the methodology used in this study to investigate the history of required skills in the Saudi labor market using machine learning. This involved job data collection and cleaning, skill extraction and labeling using NLP, time and industry trend analysis, machine learning model measurement, and comparison against educational programs.

In spite of the limitations, the method worked effectively in narrowing down key skills trends and shortfalls. The findings will be the basis for the next chapter wherein discussion and conclusions will be extensively outlined.