King Saud University
College of Computer and Information Sciences
Department of Information Technology
IT 362: Principles of Data Science
Project – Phase#1

saudi law
نظام العمل

444200965	ماريه صالح النفيسه
444200498	رغد نوري الرشيد
444200652	لين خالد الدبيس
444204601	ملاك طلال باسلوم
444201090	نوره صالح الخضير

Introduction:

The Saudi Labor Law is considered one of the most important regulations that ensure a fair and motivating work environment for all employees in the Kingdom, as it aims to protect workers' rights and regulate the relationship between employees and employers, thereby promoting workplace justice. Despite its importance, this law consists of a set of legal rules written in specialized language and filled with technical terms. Here lies the problem: the difficulty of quickly accessing the required information, as non-specialists in law—such as employees, employers, and others—face major challenges when searching for a specific legal article in a short time.

Therefore, this project aims to transform the articles of the Saudi Labor Law into a structured Question–Answer (Q&A) dataset that enables various beneficiary groups to quickly access legal texts. This highlights the importance of presenting these legal materials in a way that is clear, organized, and easily searchable for all users.

Research Question: How can the articles of the Saudi Labor Law be transformed into a structured Question–Answer dataset that supports quick access to legal information and serves various beneficiary groups?

Data sources:

1- Bureau of Experts at the Council of Ministers
2- Istitlaa Platform
3- Qiwa Platform
4- Master's Theses from King Saud University Digital Library

Source1:

Description: The official text of the Saudi Labor Law, collected from the official website of the Bureau of Experts at the Council of Ministers. This text represents the legal framework that regulates the relationship between employees and employers in the Kingdom of Saudi Arabia.

Data Type: Unstructured text data, as the file consists of lengthy legal articles in Arabic provided in PDF format.

Potential Biases:
1-Representation: The law formally covers workers and employers, but it may not fully reflect the situations of informal or undocumented workers.

2-Measurement: The text is accurate since it is taken directly from the official source, but the specialized legal language may make access to information slower or more difficult for non-specialists.

3-Historical Bias: The text reflects only the most recent amendments and does not include previous versions of the law or rights that may have changed over time.

Sorce-link: https://www.boe.gov.sa/

Source2:

Description: Draft amendments to the Implementing Regulations of the Saudi Labor Law, published on the Istitlaa (Public Consultation Platform). The document was prepared and issued by the Ministry of Human Resources and Social Development (MHRSD) as part of Vision 2030 initiatives to enhance labor market efficiency, regulate employment practices, and gather public and stakeholder feedback before final approval.

Data Type: Semi-structured data. The core file is provided in Arabic in PDF format with regulatory articles and proposed amendments. In addition, the platform itself provides structured metadata such as the consultation period, target groups, responsible authority (MHRSD), and the project status (active/closed).

Potential Biases:
1- Representation: The draft reflects the Ministry's official proposal, but it may not represent the final legally binding regulation. It also primarily addresses the formal labor market and may not cover informal or undocumented workers.

2- Measurement: The content is accurate as it comes directly from the Ministry, but the technical and legal language may pose accessibility challenges for non-specialists.

3- Historical Bias: The draft captures only the most recent proposed changes and does not include previous versions of the regulation or a full record of earlier labor law practices.

Source-link: https://istitlaa.ncc.gov.sa/

Source3:

Description: Qiwa is the official digital platform under the Ministry of Human Resources and Social Development (MHRSD) that provides integrated electronic services for employers and workers in the private sector. The platform includes access to labor law content such as the main chapters of the Saudi Labor Law, practical guidelines, and Q&A on its implementation, in addition to automated services like employment contracts, company regulations, and work permits.

Data Type: Semi-structured text data. While the labor law is provided as official structured chapters and articles, the platform also offers practical interpretations in the form of Q&A and interactive guidance. In addition, metadata such as service categories, target users (workers/employers), and update timestamps are included.

Potential Biases:
1- Representation: The platform mainly focuses on the private sector and on users registered in the system; it may not fully reflect the realities of informal workers or sectors not covered by Qiwa services.

2- Measurement: The data is authoritative since it is maintained by MHRSD, but some of the content is presented as practical Q&A, which may simplify or omit complex legal nuances.

3- Historical Bias: The platform reflects the current law and updated practices. Older versions of the law or outdated interpretations are not preserved or displayed, which may limit longitudinal analysis.

Source-link: https://qiwa.sa/

Source4:

Description: Master's theses containing analytical studies on Saudi labor law, such as termination rules, training contracts, and remote work contracts.

Data Type: Text data, unstructured because the theses presents content as pages of paragraphs.

Title Of These Two Master's Theses:
1- "Provisions of Remote Work Contracts under the Saudi Labor Law" By: Jana Khaled Mohamed Al-Harbi & Dr. Amal Mohamed Shalaby Khodraj (2025).

2- "Termination Provisions in Training and Qualification Contracts under the Saudi Labor Law" By: Taqiyah Ali Rashwan & Dr. Al-Hanouf Al-Salmi (2024).

Potential Biases:
1- **Representation:** Covers academic legal cases; may not represent daily practical cases for all groups.

2- **Measurement:** Accurate information but analytical and unofficial compared to government sources.

3- **Historical Bias:** Relatively recent, but limited to the academic study and selected research areas.

Source: PDF Files.

Source	Observations	Features	Data Types
Bureau of Experts at the Council of Ministers	245	Section Chapter Article Text Status Source	String String String Text String String
Istitlaa Platform	16	Article Current_Text Proposed_Text URL	String text text String
Qiwa	245	URL Content	String Text
Master's Theses from King Saud University Digital Library	2	Thesis_Title Author_Name Year	String String Integer

Objectives:

The main objective of this project is to develop a structured Question–Answer (Q&A) dataset from the Saudi Labor Law that can generate meaningful insights and support practical legal tasks. This dataset will provide students, researchers, and practitioners with a simplified Q&A format that improves comprehension and facilitates effective legal review

In addition, the dataset will enable thematic analysis of the law's content, such as identifying the focus of articles on contracts, wages, leaves, termination, and disciplinary actions. It will also serve as a foundation for AI applications, supporting the training of language models to create legal assistants and intelligent search systems.

Moreover, the dataset can assist HR managers and legal professionals in addressing real-world questions—for example, probation periods, severance pay, or working hours—by linking each answer directly to official articles. Finally, it will promote transparency and compliance by ensuring that every answer remains grounded in the official legislative text, guaranteeing both accuracy and reliability.

Method:

We will begin by collecting the official Saudi Labor Law text from Qiwa.sa, the Ministry of Human Resources and master's theses from King Saud University Digital Library websites using Selenium, since these sources provide the most reliable and up to date version of the Saudi Labor law. The raw text gathered from these sources will be saved in a CSV file containing the extracted content.

Once the data is collected, we will segment the labor law into individual articles or sections to maintain clarity and preserve the original legal structure. At this stage, we will also design a clear Q&A format.

To transform the raw legal text into a Q&A dataset, we will use large language models (LLMs) to generate candidate questions and answers for each article. The focus will be on practical themes such as contracts, wages, probation periods, termination, leaves, and disciplinary actions, since these are the topics most relevant for both legal professionals and students. Importantly, we will ensure that the answers produced by the LLM remain grounded in the official legislative text so that accuracy and reliability are not compromised.

After generating the Q&A pairs, we will validate the results through manual review. This step will allow us to confirm the legal accuracy of the dataset and correct or remove any outputs where the model has misinterpreted the law.

Finally, we will consolidate the validated Q&A pairs into a structured dataset that can serve multiple purposes: as a study aid for students, as a resource for thematic analysis of Saudi labor law, and as a foundation for building AI applications such as legal assistants or intelligent search systems.

Challenges in data collection and recommendations :

Challenges in Data Collection:

- Website Structure: The Qiwa platform presents the Saudi Labor Law in separate sections under different URLs, which made the collection process more complex since the data was not centralized on a single page.
- Different Data Formats: The legal texts were straightforward, while the FAQ section required a question–answer format, creating challenges in unifying the data structure.
- WebDriver Configuration: Running Selenium in headless mode required fine-tuning to ensure proper handling of dynamic content loading.
- Web Scraping Prevention: Some websites hinder or block automated collection through dynamic JavaScript loading, protections like CAPTCHA, robots.txt restrictions, or automated behavior detection systems (rate-limiting, WAF). This caused interruptions, errors, and the risk of account or IP blocking.
- Text Encoding: Arabic text appeared corrupted when opened in Excel due to encoding issues.

Recommendations for Future Data Collection and Analysis:

- Organized CSV Files: Link each section URL with its extracted text to ensure data traceability and organization.
- Separate Data Formats: Store FAQ data in a dedicated CSV file (question–answer format) to maintain consistency.
- Enhanced WebDriver Settings: Configure Selenium options to achieve stable performance with dynamic content.
- Legal Compliance: Rely on sources that allow web scraping while ensuring permissions and compliance with regulations.
- Proper Encoding: Save files using UTF-8-SIG encoding to prevent issues with displaying Arabic text.

Jupyter Notebook :

https://colab.research.google.com/drive/1tNbPowYsRrsnrDxaYjgJ-6rAb_Nc7_qq?usp=sharingt

Log book :

https://drive.google.com/file/d/12aobyxVYAHYThpuY4Oct9YCpY796UErX/view?usp=drivesdk