PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA
MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH
HAMMA LAKHDAR ELOUED UNIVERSITY
— FACULTY OF EXACT SCIENCES —
— COMPUTER SCIENCE DEPARTMENT —

# Memory

In Candidacy for the Degree of

**Academic Licence**

OPTION : COMPUTER SCIENCE

---

# Design and develop a dataset platform for AI proposes with an Arabic oriented content

---

**Students:**

- **Arwa MEISSA**
- **Nour Houda SAIDI**
- **Ikram YOUMBAI**

**Supervisor:**
Dr. Mohammed Charaf Eddine MEFTAH
Department of Computer Science
El Oued University

El Oued, ALGERIA. May 2024

# DEDICATION

This humble effort is dedicated to A special feeling of gratitude to my loving parents, whose have taught me to work hard for the things that I aspire to achieve and their words of encouragement for tenacity still ringing in my ears.

To my loving sister **Marwa** You are not just a sister, but a soulmate, a teacher, and my role model in life. Your unwavering support, wisdom, and kindness shaped me in ways I can never fully express. It was your passion for artificial intelligence, your relentless pursuit of knowledge, and your remarkable achievements, including earning a Ph.D. in this field, that inspired me to follow the path of computer science. Your guidance and encouragement continue to guide me.

To my dear brothers and my sister in law, Your support, love, and camaraderie have been my guiding light.

This memory is dedicated to each of you for shaping who I am today.

**Arwa**

# Dedication

This memory is dedicated to **my parents**, whose unwavering support and sacrifices have been the cornerstone of my academic journey. Your love and encouragement have been my greatest motivation. To my sister **Rabia** and my brother **Ali** and **Zakaria** who have supported me throughout the process and have never left my side and are very special. To my supervisor, **Dr. MEFTAH Mohammed Charaf Eddine**, whose wisdom, guidance, and patience have been invaluable throughout this process. Your belief in my abilities has inspired me to reach new heights.

**Ikram**

# DEDICATION

Praise be to **Allah** for the joy of accomplishment, and praise be to Allah at the beginning and at the end...

To **my father**, who illuminated my paths and my way, and my role model in every step I take. To **my compassionate mother**, the warm embrace and my sky that has never abandoned me, and without whom my day is never complete. To **my brothers**, who have always stood by me and supported me throughout my educational journey. To all **my teachers** who have taught, guided, and directed me, even with a single letter. And to my friends **Arwa** and **Ikram**, I pray that success always accompanies you, dear friends with beautiful hearts and souls. Congratulations on your graduation.

I dedicate this humble work to all of you...

**Houda**

# ACKNOWLEDGEMENT

In the name of **Allah**, the most gracious, the most merciful. All thanks to **Allah**, the lord of both worlds, and may prayers and peace be upon our prophet, **Muhammad** and, upon his companions.

The past few years have been an incredible journey. This page is about the people that have supported us throughout this journey.
First of all, we would like to thank Professor MEFTAH Mohamed Charaf Eddine , our supervisor, for or their invaluable guidance, support, and encouragement throughout the course of this thesis. Their expertise and insights have been instrumental in shaping this work, and their constant feedback has been vital in refining our research.
We extend our heartfelt thanks to our teachers, whose knowledge and dedication have inspired and motivated us throughout our academic journey. Their teachings have provided a solid foundation for our research and have greatly contributed to our intellectual growth.
Lastly, We would like to thank our families and friends for their unwavering support and understanding. Their patience and encouragement have been a source of strength during challenging times.

This thesis is the result of collective efforts, and we are truly thankful to everyone who has contributed in any way to the completion of this work.

# ABSTRACT

This work investigates the critical intersection of web-based systems, artificial intelligence (AI), and datasets, with a specific emphasis on addressing the scarcity of Arabic-oriented content in AI applications. The study begins by elucidating preliminary notions, exploring web-based systems, AI, datasets, and the unique challenges posed by Arabic content. A comprehensive literature review underscores the importance of datasets in AI development and the pressing need for Arabic-oriented content platforms. The Related Works section examines existing dataset platforms, particularly those catering to Arabic content, and classifies them by region. The Planning phase outlines the objectives, requirements, and business model for the proposed project. The Conception phase delves into the Front-end and Back-end conceptualization processes, providing detailed insights into sitemaps, wireframes, platform architecture, and database description. The Implementation and Development phase elaborates on the tools utilized and the process of developing DataBaazar, a platform aimed at bridging the gap in Arabic-oriented content for AI applications. Major findings highlight the necessity of such platforms to foster AI development in Arabic-speaking regions and the challenges in identifying and classifying relevant datasets. The study underscores the importance of user experience and interface design in enhancing the accessibility and usability of dataset platforms. In conclusion, this thesis contributes to advancing AI development by proposing a solution to the scarcity of Arabic-oriented content, thereby facilitating the creation of more inclusive and diverse AI applications tailored to Arabic-speaking populations.



**Keywords:**
Data set, Platform, Arabic Content, Artificial Intelligence, .



## الملخص:

يبحث هذا العمل في التقاطع الحاسم بين الأنظمة القائمة على الويب والذكاء الاصطناعي ومجموعات البيانات، مع التركيز بشكل خاص على معالجة ندرة المحتوى الموجه نحو اللغة العربية في تطبيقات الذكاء الاصطناعي. تبدأ الدراسة بتوضيح المفاهيم الأولية، واستكشاف الأنظمة القائمة على الويب، والذكاء الاصطناعي، ومجموعات البيانات، والتحديات الفريدة التي يطرحها المحتوى العربي. تؤكد مراجعة شاملة للأدبيات على أهمية مجموعات البيانات في تطوير الذكاء الاصطناعي والحاجة الملحة لمنصات المحتوى ذات التوجه العربي. ويبحث قسم الأعمال ذات الصلة منصات مجموعة البيانات القائمة، ولا سيما تلك التي تلبي المحتوى العربي، ويصنفها حسب المنطقة. وتحدد مرحلة التخطيط أهداف المشروع المقترح ومتطلباته ونموذج أعماله. تتعمق مرحلة الحمل في عمليات التصور الأمامي والخلفي، مما يوفر رؤى مفصلة في خرائط المواقع والإطارات اللاسلكية وهندسة المنصة ووصف قاعدة البيانات. وتتناول مرحلة التنفيذ والتطوير بالتفصيل الأدوات المستخدمة وعملية تطوير DataBaazar، وهي منصة تهدف إلى سد الفجوة في المحتوى الموجه نحو اللغة العربية لتطبيقات الذكاء الاصطناعي. تسلط النتائج الرئيسية الضوء على ضرورة هذه المنصات لتعزيز تطوير الذكاء الاصطناعي في المناطق الناطقة باللغة العربية والتحديات في تحديد وتصنيف مجموعات البيانات ذات الصلة. تؤكد الدراسة على أهمية تجربة المستخدم وتصميم الواجهة في تعزيز إمكانية الوصول إلى منصات مجموعة البيانات وإمكانية استخدامها. وفي الختام، تساهم هذه الأطروحة في النهوض بتنمية الذكاء الاصطناعي من خلال اقتراح حل لندرة المحتوى الموجه نحو اللغة العربية، مما يسهل إنشاء تطبيقات أكثر شمولاً وتنوعًا للذكاء الاصطناعي مصممة خصيصًا للسكان الناطقين باللغة العربية.



**الكلمات المفتاحية:** مجموعة البيانات، المنصة، المحتوى العربي، الذكاء الاصطناعي

# Introduction

The rapid advancement of artificial intelligence (AI) technologies has revolutionized various sectors, driving the need for high-quality datasets to fuel the development and refinement of AI models. Datasets serve as the backbone for machine learning algorithms, providing the necessary information for training and validation. However, the availability and quality of these datasets vary significantly across different regions and languages, posing unique challenges and opportunities.

This thesis explores the current state of datasets for AI applications, with a particular focus on Arabic content. While substantial progress has been made in creating and curating datasets in widely spoken languages like English, there remains a significant gap in resources for Arabic, a language spoken by over 400 million people worldwide. The lack of comprehensive and accessible Arabic datasets hampers the development of AI applications tailored to Arabic-speaking users, limiting the potential benefits of AI in these regions.

In addressing this gap, the thesis is structured as follows: Chapter 1 provides an overview of preliminary notions related to web-based systems, artificial intelligence, and the importance of datasets. It also emphasizes the significance of Arabic content in the digital age. Chapter 2 reviews related works, classifying dataset platforms by region and highlighting the challenges associated with identifying platforms for Arabic content. Subsequent chapters detail the objectives, requirements, and content strategy necessary for developing a robust dataset platform for Arabic AI applications. The conception phase, discussed in Chapter 4, outlines the front-end and back-end conceptualization, including sitemaps, wireframes, and database diagrams. Finally, Chapter 5 covers the implementation and development processes, focusing on the tools and methodologies employed to create the DataBaazar platform, aimed at enhancing the accessibility and quality of Arabic datasets for AI applications.

Through this structured approach, the thesis aims to contribute to the body of knowledge in AI development and highlight the critical need for Arabic-oriented datasets. By addressing the existing challenges and proposing practical solutions, this research endeavors to bridge the gap in AI resources, fostering innovation and inclusivity in the rapidly evolving digital landscape.

# List of Figures

**Figure** **Page**

# List of Tables

**Table** **Page**

# TABLE OF CONTENTS

CHAPTER 1

# STATE OF THE ART

## 1.1 Introduction

In the rapidly advancing landscape of technology, understanding fundamental concepts is paramount to delving deeper into specialized areas. This chapter serves as a primer, exploring essential preliminary notions essential to comprehend the subsequent discourse. From the ubiquitous nature of web-based systems to the transformative power of artificial intelligence (AI), each concept elucidates a crucial aspect of modern computing. Additionally, datasets stand as the lifeblood of AI, shaping its capabilities and applications. Moreover, the richness of Arabic content within these datasets highlights the importance of cultural diversity in technological innovation. By exploring these foundational concepts, we lay a solid groundwork for the exploration of more complex topics in subsequent chapters.

## 1.2 Preliminary notions

### 1.2.1 Web-based systems

Web-based systems are software applications accessed through web browsers, offering accessibility from any device with internet connectivity. They boast platform independence, functioning seamlessly across various operating systems. Centralized management facilitates updates and maintenance, ensuring consistency. Additionally, they exhibit

scalability, accommodating changing user demands and system requirements efficiently.

#### 1.2.1.1 Definition

A web-based system is a software application or platform that operates over the internet and is accessed through a web browser. Unlike traditional software that is installed locally on a user's device, web-based systems are hosted on remote servers and delivered to users via the internet. Users interact with web-based systems through a graphical user interface (GUI) presented in their web browser, allowing them to access and utilize the system's features and functionalities from any device with internet connectivity. Web-based systems can range from simple applications, such as online calculators or collaboration tools, to complex enterprise-level systems, such as customer relationship management (CRM) software, project management tools, or online banking platforms. These systems often involve database storage, user authentication, data processing, and other functionalities beyond simply displaying information.

#### 1.2.1.2 Characteristics

Key characteristics of web-based systems include:

- Accessibility: Users can access the system from anywhere with an internet connection, using a variety of devices such as desktop computers, laptops, tablets, and smartphones.
- Platform Independence: Web-based systems are not tied to any specific operating system or device platform. They can be accessed and used on different devices and operating systems without requiring separate versions or installations.
- Centralized Management: Since the software is hosted on remote servers, updates, maintenance, and data management are typically centralized, reducing the burden on individual users and ensuring consistency across the system.
- Scalability: Web-based systems can easily scale to accommodate changes in user demand or system requirements by adding more server resources or optimizing performance.
- Collaboration: Many web-based systems support collaborative features, allowing multiple users to work together in real-time, share data, and communicate within the system.

### 1.2.2 Artificial intelligence

Artificial Intelligence (AI) has undergone significant evolution, profoundly influencing diverse facets of human existence. This technology, while captivating, has brought about substantial changes in our lifestyles, professions, and interactions with our environment. Across the expansive realm of AI, numerous specialized domains emerge, each possessing distinct attributes and practical uses.

#### 1.2.2.1 Definition

The definition of Artificial Intelligence (AI) encompasses the simulation of human intelligence processes by machines, particularly computer systems. It involves machines performing tasks that typically require human intelligence, such as expert systems, natural language processing, speech recognition, and machine vision [57]. AI can be categorized into weak AI, which is task-specific and designed for a particular function like virtual personal assistants, and strong AI, also known as artificial general intelligence (AGI), which aims to replicate the cognitive abilities of the human brain and can adapt to unfamiliar tasks autonomously. John McCarthy, an emeritus Stanford Professor, coined the term AI in 1955, defining it as "the science and engineering of making intelligent machines" . The definition of AI has evolved to emphasize machines that can learn, akin to human learning processes, and autonomous systems that can independently plan and decide steps to achieve goals without constant human intervention

#### 1.2.2.2 AI Applications

- Machine Learning (ML): Machine Learning focuses on developing complex algorithms and models that allow machines to learn from a set of data and make predictions or decisions. It finds applications in various fields, such as healthcare (for diagnosis and prognosis), finance (for fraud detection), and Natural Language Processing (for chatbots and language translation). ML serves as the foundation for many other AI domains [54].

- Natural Language Processing (NLP): NLP enables computers to understand, interpret, and generate human language. It is crucial for chatbots, language translation, sentiment analysis, and speech recognition systems [43].

- Computer Vision: Computer vision involves teaching machines to interpret and analyze visual information from images or videos. It has applications in fields like

autonomous vehicles, medical imaging, and surveillance [32].

- Robotics: Robotics combines AI, sensors, and mechanical engineering to create intelligent machines capable of performing tasks autonomously. It plays a significant role in manufacturing, healthcare, and exploration [34].
- Expert Systems: Expert systems mimic human expertise by using knowledge-based rules to solve complex problems. They find applications in fields like medical diagnosis and decision support [58].
- Autonomous Systems: Autonomous systems operate independently, making decisions without human intervention. Examples include self-driving cars and drones [52]

### 1.2.3 Datasets

#### 1.2.3.1 Definition

According Merriam Webster dictionary the dataset is defined as a collection of data taken from a single source or intended for a single project [23]. The authors in [46] defined is as a collection of data items (facts, measurements, or observations) about a set of variables. Each data item is called a record, and it is typically represented by a row in a table. Each variable is represented by a column in the table. In simpler terms, A dataset is a structured collection of data, typically organized in a tabular format with rows and columns, or sometimes in other formats like JSON, XML, or relational databases. It represents a coherent set of information relevant to a particular domain, problem, or analysis. Datasets can be small or large, ranging from a few data points to millions or even billions of records, depending on the context. They are commonly used in fields such as statistics, machine learning, data analysis, and data science for various purposes like training models, testing hypotheses, or gaining insights from the data.

#### 1.2.3.2 Dataset types

There are many different types of datasets, each with its own unique characteristics and uses. Here are some of the most common types:

- **Numerical datasets:** These datasets contain data that can be represented by numbers. Examples of numerical data include temperature, humidity, sales figures, and stock prices [47].

| Objects | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ | $a_6$ | Decision |
|---|---|---|---|---|---|---|---|
| $x_1$ | 1 | 7 | 1 | 1 | 1 | 2 | 2 |
| $x_2$ | 2 | 1 | 5 | 1 | 1 | 5 | 1 |
| $x_3$ | 1 | 3 | 4 | 1 | 1 | 4 | 2 |
| $x_4$ | 2 | 2 | 3 | 3 | 1 | 4 | 1 |
| $x_5$ | 1 | 3 | 1 | 1 | 1 | 2 | 2 |
| $x_6$ | 2 | 1 | 4 | 1 | 2 | 4 | 1 |
| $x_7$ | – | – | 2 | 2 | 2 | 1 | 2 |
| $x_8$ | 1 | 3 | 2 | 3 | 1 | 2 | 2 |
| $x_9$ | 1 | 8 | 4 | 1 | 2 | 5 | 1 |
| $x_{10}$ | – | – | – | 1 | 2 | – | – |
| $x_{11}$ | – | 3 | – | – | – | 2 | 1 |
| $x_{12}$ | 1 | 2 | 4 | 3 | 2 | 4 | 1 |

Figure 1.1: Example of a Numerical dataset

- **Categorical datasets:** These datasets contain data that can be classified into categories. Examples of categorical data include gender, hair color, occupation, and blood type [36].

| Color |
|---|
| Red |
| Red |
| Yellow |
| Green |
| Yellow |

| Red | Yellow | Green |
|---|---|---|
| 1 | 0 | 0 |
| 1 | 0 | 0 |
| 0 | 1 | 0 |
| 0 | 0 | 1 |
| 0 | 1 | 0 |

Figure 1.2: Example of a Categorical dataset

- **Text datasets:** These datasets contain textual data, such as words, sentences, or documents. Text datasets are often used in natural language processing (NLP) tasks, such as sentiment analysis and machine translation [55].

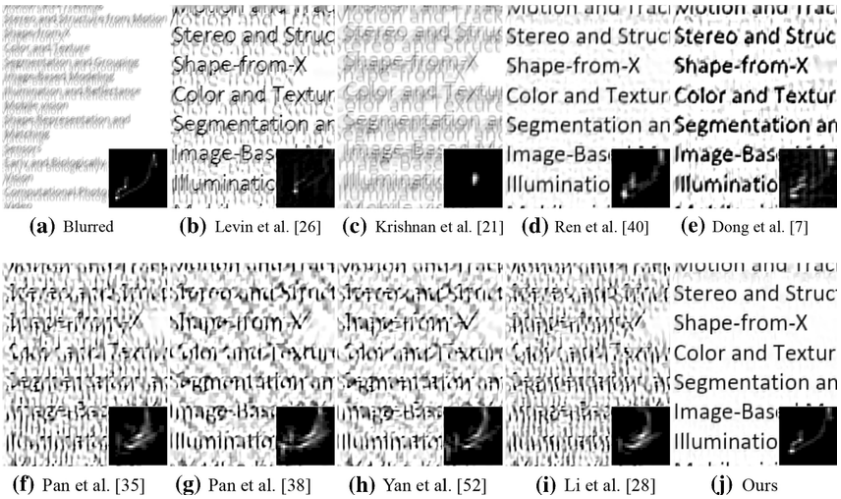


Figure 1.3: Example of a Text dataset

- **Image datasets:** These datasets contain images, such as photographs, satellite imagery, or medical scans. Image datasets are often used in computer vision tasks, such as object recognition and image classification [51].

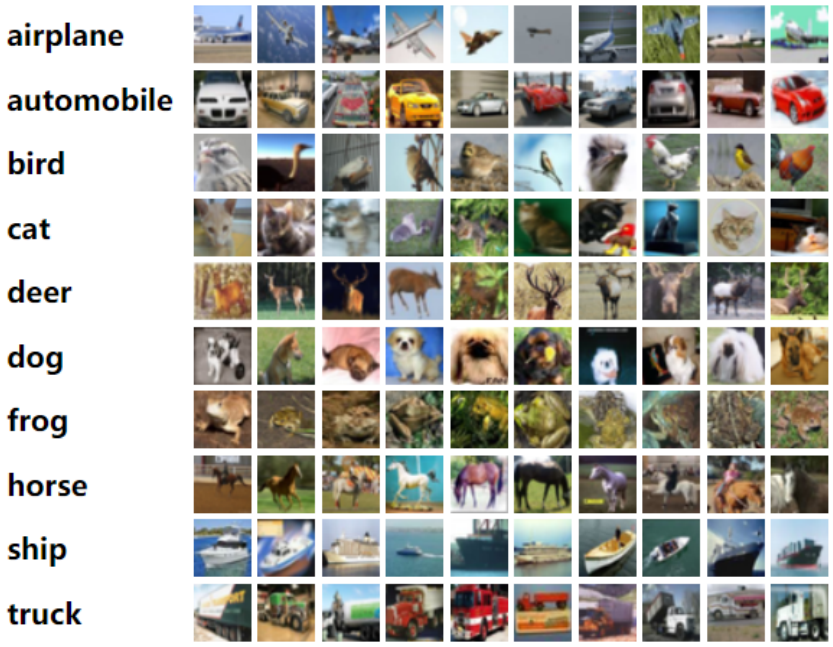


Figure 1.4: Example of an Image dataset

- **Audio datasets:** These datasets contain audio recordings, such as speech, music, or environmental sounds. Audio datasets are often used in speech recognition and

speaker identification tasks [49].

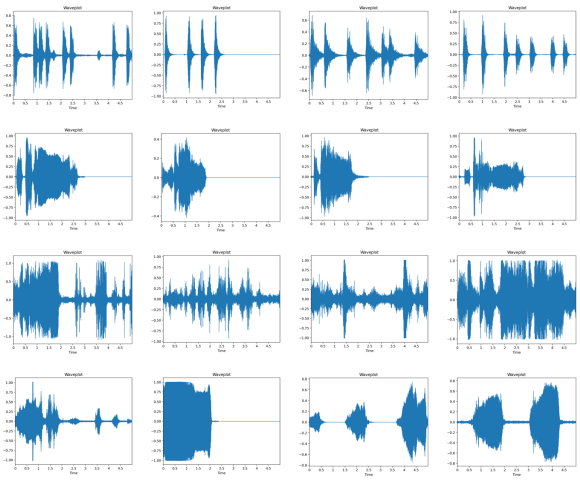

Figure 1.5: Example of an Audio dataset

- **Video datasets:** These datasets contain videos, which are collections of images played back over time. Video datasets are often used in tasks such as action recognition and video summarizing [33].

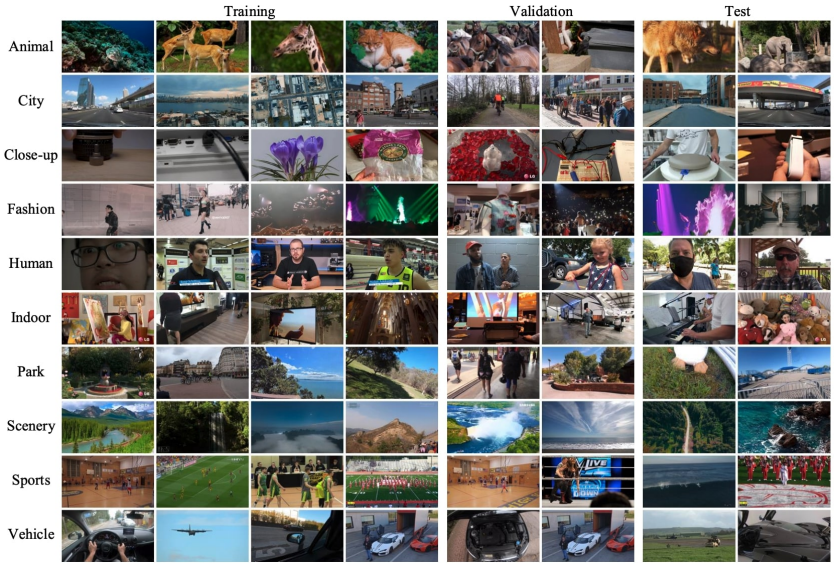


Figure 1.6: Example of a Video dataset

- **Time series datasets:** These datasets contain data that is collected over time. Examples of time series data include stock prices, weather data, and website traffic [41].

| | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | id | name | latitude | longitude | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 |
| 2 | 1 | Atlanta | 33.7489 | -84.3881 | 85 | 38 | 75 | 30 | 9 | 15 | 38 |
| 3 | 2 | Chicago | 41.85 | -87.65 | 28 | 29 | 38 | 26 | 15 | 12 | 10 |
| 4 | 3 | Dallas | 32.7828 | -96.8039 | 18 | 59 | 22 | 60 | 82 | 42 | 18 |
| 5 | 4 | Denver | 39.7392 | -104.9842 | 35 | 45 | 31 | 26 | 14 | 9 | 15 |
| 6 | 5 | Houston | 29.7631 | -95.3631 | 12 | 31 | 15 | 22 | 28 | 38 | 31 |
| 7 | 6 | Kansas City | 39.0997 | -94.5783 | 25 | 50 | 25 | 25 | 25 | 25 | 100 |
| 8 | 7 | Los Angeles | 34.0522 | -118.2428 | 88 | 46 | 56 | 15 | 12 | 25 | 46 |
| 9 | 8 | Miami | 25.7738 | -80.1924 | 52 | 51 | 46 | 68 | 75 | 85 | 96 |
| 10 | 9 | Minneapolis | 44.98 | -93.2636 | 7 | 12 | 18 | 11 | 9 | 9 | 4 |
| 11 | 10 | New York | 40.7142 | -74.0064 | 23 | 18 | 16 | 24 | 26 | 28 | 30 |
| 12 | 11 | Philadelphia | 39.9522 | -75.1642 | 32 | 28 | 29 | 25 | 22 | 15 | 8 |
| 13 | 12 | Phoenix | 33.4539 | -112.0746 | 8 | 15 | 22 | 25 | 29 | 28 | 32 |
| 14 | 13 | San Francisco | 37.775 | -122.4183 | 82 | 74 | 72 | 10 | 85 | 88 | 74 |
| 15 | 14 | Seattle | 47.6097 | -122.3331 | 9 | 16 | 14 | 23 | 45 | 66 | 85 |
| 16 | 15 | Washington | 38.89 | -77.03 | 33 | 45 | 68 | 96 | 102 | 82 | 74 |
| 17 | | | | | | | | | | | |

Figure 1.7: Example of Time series dataset

- **Relational datasets:** These datasets are stored in relational databases, which are a type of database that allows you to store and access data in a structured way. Relational datasets are often used in business intelligence and data warehousing applications [48].

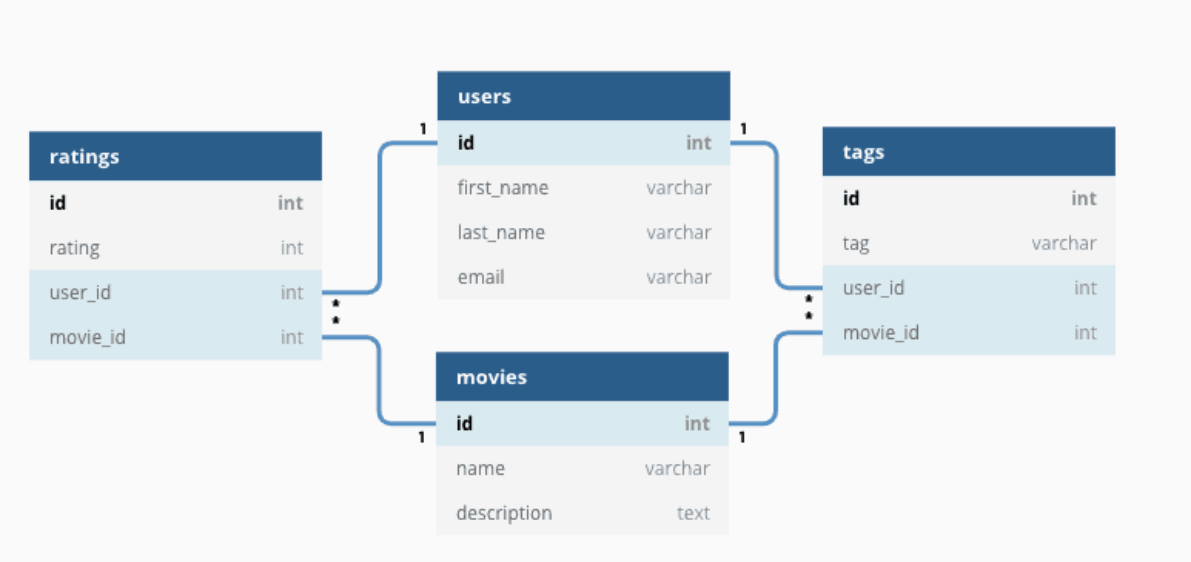


Figure 1.8: Example of a relational dataset

### 1.2.4 Arabic content

Arabic content refers to any information, text, media, or material that is written or produced in the Arabic language. This includes written texts, websites, articles, books, social media posts, videos, audio recordings, and any other form of communication created using Arabic script. Arabic content can cover a wide range of topics and subjects, including literature, religion, science, history, news, entertainment, and more. It is an essential aspect of cultural expression and communication for Arabic-speaking communities around the world.

## 1.3 Literature review

### 1.3.1 Overview of the importance of datasets in AI development

Datasets are the fundamental building blocks of Artificial Intelligence, especially in the realm of Machine Learning. They act as the fuel that powers these intelligent systems, allowing them to learn, grow, and make accurate predictions. Here's a breakdown of their importance in AI development:

- Training Data: Datasets provide the raw material used to train machine learning algorithms. By exposing AI models to large amounts of diverse and representative data, they can learn patterns, relationships, and features that enable them to make accurate predictions or classifications.
- Performance Improvement: The quality and size of the dataset directly impact the performance of AI models. Larger, more comprehensive datasets often lead to better-performing models with higher accuracy and generalization ability.
- Generalization and Robustness: Datasets help AI models generalize their learning beyond the examples seen during training, allowing them to make accurate predictions on unseen data. Well-curated datasets can also enhance the robustness of AI systems by exposing them to various edge cases and scenarios.
- Bias Mitigation: Datasets can inadvertently contain biases present in the data collection process or in society. Recognizing and addressing biases in datasets is essential for developing fair and ethical AI systems that do not discriminate against certain groups or perpetuate societal inequalities.

- Domain-Specific Knowledge: Datasets capture domain-specific knowledge and expertise, enabling AI systems to perform tasks and make decisions within specific domains such as healthcare, finance, natural language processing, computer vision, and more.
- Innovation and Research: Access to high-quality datasets fuels innovation and advances in AI research. Researchers rely on publicly available datasets to benchmark algorithms, replicate results, and develop new techniques and approaches in machine learning.
- Customization and Personalization: Datasets can be tailored to specific use cases or applications, allowing developers to fine-tune AI models for particular tasks or user preferences. Customized datasets enable personalized recommendations, content filtering, and adaptive user interfaces.
- Data Privacy and Security: Proper management of datasets is critical for ensuring data privacy and security. AI developers must adhere to regulations and best practices for handling sensitive or personal information, including data anonymization, encryption, and access control measures.
- Continual Learning and Adaptation: Datasets are not static; they evolve over time as new data becomes available or as the underlying distribution of the data changes. AI systems must be capable of continual learning and adaptation to remain effective in dynamic environments.
- Real-World Applications: Ultimately, the value of AI lies in its ability to solve real-world problems and address societal challenges. Datasets enable the development of AI applications that have tangible benefits across various domains, including healthcare, transportation, finance, agriculture, and more.

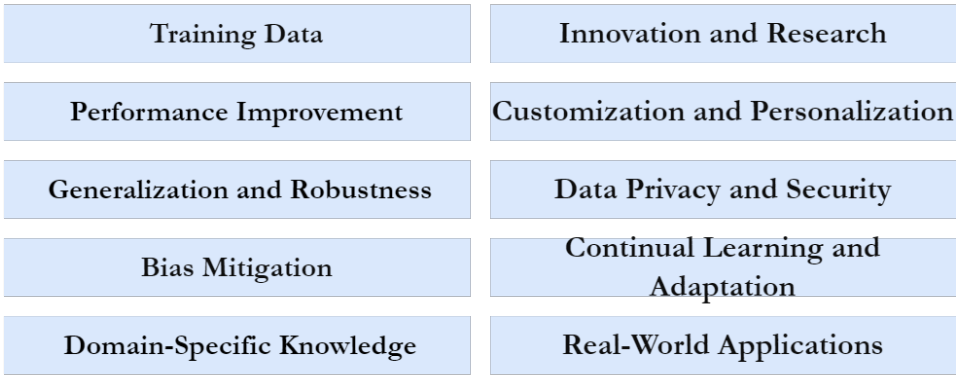


Figure 1.9: The importance of datasets in AI applications

### 1.3.2 Arabic oriented content in AI applications

Arabic-oriented content in AI applications refers to the adaptation of artificial intelligence technologies to better cater to Arabic-speaking users, their language, culture, and specific needs. This can include various aspects:

- Language Processing: AI models need to be trained on large amounts of Arabic text to understand and process the language effectively. This involves Natural Language Processing (NLP) tasks such as text analysis, sentiment analysis, and machine translation [45].
- Speech Recognition and Synthesis: Developing accurate Arabic speech recognition and synthesis systems is crucial for applications like virtual assistants, speech-to-text systems, and voice-controlled devices [56].
- Language Models: Building Arabic language models like GPT (Generative Pre-trained Transformer) models helps in generating human-like text, understanding context, and providing relevant responses in Arabic [38].
- Localization: Adapting AI interfaces, applications, and content to the Arabic language and culture. This involves not only translating text but also considering cultural nuances, localizing graphics, and ensuring usability for Arabic speakers [53].
- Content Recommendation: Recommender systems tailored to Arabic users' preferences and interests, considering cultural factors and language-specific content [35].

- Chatbots and Virtual Assistants: Developing conversational AI agents that can understand and respond effectively in Arabic, offering support, information, and assistance in various domains [40].
- Search Engines: Optimizing search engines to provide accurate and relevant results for Arabic language queries, understanding Arabic semantics, and handling dialect variations [29].
- E-learning and Education: Providing educational content in Arabic, including online courses, tutorials, and interactive learning platforms, to support Arabic-speaking learners [30].
- Healthcare Applications: Developing AI-driven healthcare applications that cater to Arabic-speaking populations, providing medical information, telemedicine services, and health advice in Arabic [31].
- Government Services: Implementing AI solutions for government services and administration in Arabic-speaking regions, such as chatbots for citizen support, Arabic language interfaces for government websites, and Arabic content for public information campaigns [59].

### 1.3.3 The importance of datasets oriented Arabic content in the digital age

In the digital age, the importance of datasets oriented towards Arabic content cannot be overstated. As Arabic-speaking populations increasingly engage with digital platforms and technologies (about 372.7 million speakers [28]), the availability of high-quality, culturally relevant datasets is essential for the development of AI systems tailored to Arabic language and culture. These datasets serve as the building blocks for training AI models across a wide range of applications, from natural language processing and sentiment analysis to image recognition and recommendation systems. By harnessing datasets that accurately reflect the linguistic nuances, cultural context, and societal norms of Arabic-speaking communities, developers can ensure that AI technologies are not only linguistically accurate but also culturally sensitive and inclusive. Moreover, the development of Arabic-oriented datasets fosters innovation and research within the Arabic-speaking world, empowering local researchers, entrepreneurs, and innovators to leverage AI for addressing pressing societal challenges, promoting economic growth,

and preserving Arabic language and culture in the digital landscape. Additionally, these datasets play a crucial role in bridging the digital divide by enabling the creation of AI-driven applications and services that cater to the needs and preferences of Arabic-speaking users, thereby promoting digital inclusion and empowerment across the region. Furthermore, as AI continues to reshape various sectors such as healthcare, education, finance, and governance, the availability of datasets oriented towards Arabic content becomes increasingly vital for ensuring that AI-driven solutions are accessible, effective, and equitable for Arabic-speaking populations. In essence, datasets oriented towards Arabic content represent not only a technological necessity but also a strategic asset for advancing AI development, fostering innovation, and promoting cultural diversity and inclusion in the digital age.

## 1.4 Summary

In conclusion, this chapter has provided a comprehensive overview of the state of the art in the realm of memory, focusing on preliminary notions crucial to understanding the subsequent discussions. The exploration of websites, artificial intelligence, datasets, and Arabic content has laid the groundwork for delving deeper into the significance of dataset-oriented Arabic content in the digital age. Through a thorough literature review, we have underscored the pivotal role of datasets in AI development and highlighted the emerging importance of Arabic-oriented content in AI applications. This sets the stage for further exploration into the intricacies of memory within the context of AI and Arabic language processing, underscoring its significance in shaping the future landscape of technology and information accessibility.

CHAPTER 2

# RELATED WORKS

## 2.1 Introduction

Dataset platforms play a critical role in fueling advancements in AI. They offer a diverse array of datasets tailored for training and evaluating AI models across domains such as computer vision, natural language processing, and reinforcement learning. Understanding the regional distribution of these platforms is essential, as it sheds light on the advantages and considerations when choosing between global platforms and those focused on specific regions. Particularly in Arabic-speaking regions, there is a growing need for platforms catering specifically to Arabic content, addressing both available resources and the challenges associated with their identification and accessibility. Navigating this landscape enables researchers and developers to make informed decisions in selecting datasets, particularly when targeting AI applications in the Arabic language and its regions.

## 2.2 Dataset platforms for AI applications

In today's data-driven world, access to high-quality datasets is essential for researchers, analysts, developers, and policymakers to drive innovation, make informed decisions, and address complex challenges. Dataset platforms play a pivotal role in providing a centralized hub for accessing, sharing, and collaborating on diverse datasets across various domains and regions. This section aims to provide an overview of prominent

dataset platforms that host datasets specifically tailored for AI applications:.

- **Kaggle** [20]: Kaggle is a popular platform where data scientists and machine learning practitioners can find datasets, participate in competitions, and collaborate on projects. It hosts a diverse range of datasets suitable for AI applications, covering various domains such as computer vision, natural language processing, and reinforcement learning.
- **UCI Machine Learning Repository** [27]: The UCI Machine Learning Repository, maintained by the University of California, Irvine, offers a collection of datasets that are commonly used for research and experimentation in machine learning. These datasets cover a wide range of AI-related topics and can be utilized for tasks such as classification, regression, and clustering.
- **ImageNet** [19]: ImageNet is an image database organized according to the WordNet hierarchy. It contains millions of labeled images across thousands of categories, making it a valuable resource for training and evaluating image classification and object detection models.
- **COCO (Common Objects in Context)** [13]: COCO is a large-scale dataset designed for object detection, segmentation, and captioning tasks. It includes images annotated with object bounding boxes, object segmentation masks, and corresponding captions, providing rich data for training AI models.
- **OpenAI Gym** [25]: OpenAI Gym is a toolkit for developing and comparing reinforcement learning algorithms. It provides a variety of environments, including classic control tasks, Atari games, and robotics simulations, enabling researchers to benchmark and evaluate AI agents in different settings.
- **Google AI Datasets** [16]: Google hosts a collection of datasets curated for AI research and development. These datasets cover diverse domains such as image recognition, natural language understanding, and structured data analysis, offering valuable resources for training and testing AI models.
- **Microsoft Research Open Data** [24]: Microsoft Research Open Data provides access to a wide range of datasets contributed by Microsoft and its collaborators. These datasets encompass various domains, including computer vision, speech recognition, and healthcare, supporting research and innovation in AI.

- **Stanford Question Answering Dataset (SQuAD)** [26]: SQuAD is a benchmark dataset for machine comprehension tasks, where models are tasked with answering questions based on given context passages. It consists of real questions posed by crowdworkers on a set of Wikipedia articles, serving as a challenging testbed for natural language understanding systems.
- **Fast.ai Datasets** [14]: Fast.ai, an organization focused on making deep learning accessible, maintains a collection of datasets suitable for deep learning projects. These datasets cover a wide range of applications, from image classification to text generation, and are often used in conjunction with Fast.ai's educational materials and courses.
- **Hugging Face Datasets** [18]: Hugging Face, known for its contributions to natural language processing (NLP), offers a repository of datasets specifically curated for NLP tasks. These datasets include benchmarks, pre-trained models, and evaluation metrics, facilitating research and development in NLP and related areas.
- **GitHub** [15]: While not exclusively a dataset platform, GitHub hosts many publicly available datasets. Developers can find datasets related to various domains and collaborate on data-driven projects.
- **Google Dataset Search** [17]: Google's search engine specifically designed for discovering datasets. It aggregates datasets from various sources and provides easy access to relevant data.
- **Amazon Web Services (AWS) Public Datasets** [1]: AWS hosts a collection of large-scale datasets across various domains, including genomics, astronomy, and climate research.

The section discusses various dataset platforms tailored for AI applications, highlighting their significance in today's data-driven world (Table 2.1). These platforms serve as centralized hubs for accessing diverse datasets across different domains and regions, essential for driving innovation and making informed decisions. Prominent platforms like Kaggle, UCI Machine Learning Repository, and ImageNet offer datasets suitable for classification, regression, and computer vision tasks. Additionally, platforms like COCO and SQuAD specialize in object detection, segmentation, and natural language processing tasks, respectively. Moreover, GitHub and Google Dataset Search provide access to publicly available datasets, while Microsoft Research Open Data offers a wide

range of datasets contributed by Microsoft and its collaborators. These platforms play a crucial role in advancing AI research and development by providing researchers, analysts, developers, and policymakers with valuable resources for training and testing AI models across various domains.

| Platform | Publication year | Focus | Data types | Dataset inventory) | AI Applications |
|---|---|---|---|---|---|
| Kaggle | 2010 | General AI | Diverse (images, text, code) | Millions | Classification, regression, NLP, reinforcement learning |
| UCI Machine Learning Repository | 1991 | Machine Learning Research | Structured data (tabular) | Hundreds | Classification, regression, clustering |
| ImageNet | 2009 | Image Recognition | Images | Millions | Image classification, object detection |
| COCO (Common Objects in Context) | 2014 | Object Detection, Segmentation | Images with annotations | Hundreds of thousands | Object detection, image segmentation, image captioning |
| OpenAI Gym | 2016 | Reinforcement Learning | Environment simulations | Dozens | Reinforcement learning |
| Google AI Datasets | N/A | General AI | Diverse (depends on dataset) | Varies | Diverse (depends on dataset) |
| Microsoft Research Open Data | 2018 | General AI | Diverse (depends on dataset) | Varies | Diverse (depends on dataset) |
| SQuAD (Stanford Question Answering Dataset) | 2016 | Natural Language Processing | Text passages and questions | Multiple versions | Machine comprehension, question answering |
| Fast.ai Datasets | 2017 | Deep Learning | Diverse (depends on dataset) | Varies | Diverse (depends on dataset) |
| Hugging Face Datasets | 2016 | Natural Language Processing | Text and code | Thousands | NLP tasks (classification, translation, summarization) |
| GitHub | 2008 | General | Diverse (depends on repository) | Millions | Diverse (depends on repository) |
| Google Dataset Search | 2020 | Dataset Discovery | Links to datasets (various types) | N/A (aggregates datasets) | N/A (discovery tool) |

Table 2.1: Datasets platforms for AI applications

## 2.3 Classification of Dataset platforms by Region

In this section, we provide descriptions of dataset platforms categorized by region, offering insights into the diversity and accessibility of data resources worldwide. Dataset platforms can be broadly categorized into two types: global and regional.

- **Global dataset platforms**: such as Kaggle, GitHub, and Google Dataset Search, provide access to datasets from diverse sources worldwide without specific regional constraints. These platforms host datasets covering a wide range of topics and domains, catering to a global audience of researchers, developers, and data enthusiasts. They offer valuable resources for collaborative research, innovation, and knowledge sharing on a global scale.

- **Regional dataset platforms**: focus on datasets specific to a particular geographical region or jurisdiction. These platforms, such as national data portals, regional data networks, and academic repositories, curate datasets that are relevant to the needs and interests of a particular region or community. Regional dataset platforms often provide access to government data, socioeconomic indicators, environmental data, and other locally relevant datasets. They play a crucial role in supporting evidence-based decision-making, policy development, and socioeconomic research at the regional and local levels.

The table 2.2 compares regional and global dataset platforms for AI applications across various features. Regional platforms focus on providing data specific to a particular country or region, catering to researchers and developers working on AI applications relevant to that area. Examples include the UK government open data platform. These platforms offer deeper insights into local trends and demographics but may have limited data volume and scope compared to global platforms. They primarily use the language(s) of the specific region. In contrast, global platforms like Kaggle offer data encompassing a wide range of geographical locations, targeting researchers and developers with global AI applications in mind. They provide a broader variety of data for diverse AI applications, though finding regionally relevant data may require additional effort. Global platforms primarily use English, with some multilingual datasets available. Examples of data from regional platforms include traffic patterns in a specific city (US Taxi Trip Data), while global platforms may offer satellite imagery of the entire Earth (EarthData). AI applications developed using regional data might include crime prediction models for a specific city (US), whereas global platforms might focus on image recognition models for

global applications, such as facial recognition. Overall, the choice between regional and global dataset platforms depends on the specific needs and scope of the AI project.

| Feature | Regional Platforms | Global Platforms |
|---|---|---|
| Focus | Data specific to a particular country or region | Data encompassing a wide range of geographical locations |
| Target Audience | Researchers and developers working on AI applications relevant to the region | Researchers and developers with global AI applications in mind |
| Examples | UK government open data (UK) | Kaggle (Global) |
| Benefits | Provides deeper insights into local trends and demographics | Offers a broader variety of data for diverse AI applications |
| Drawbacks | Limited data volume and scope | May require additional effort to find regionally relevant data |
| Languages | Primarily language(s) of the specific region | Primarily English, with some multilingual datasets |
| Examples of Data | Traffic patterns in a specific city (US Taxi Trip Data) | Satellite imagery of the entire Earth (EarthData) |
| Examples of AI Applications | Crime prediction models for a specific city (US) | Image recognition models for global applications (facial recognition) |

Table 2.2: Regional vs. Global Datasets platforms for AI applications

## 2.4 Arabic oriented content Dataset Platforms for AI Applications

### 2.4.1 Arabic oriented content Dataset Platforms

Identifying specific dataset platforms dedicated solely to the Arabic world and Arabic content can be difficult. Many global platforms might host Arabic datasets, but they wouldn't be the primary focus. However, there are still ways to find valuable resources:

- **University Repositories:** Universities in Arabic-speaking countries often maintain repositories containing datasets relevant to their research areas. Look for universities with strong computer science or NLP programs and explore their online resources.

- **Arabic NLP Initiatives:** Organizations dedicated to Arabic NLP development might curate or host Arabic-specific datasets [39]. Search for initiatives like the Arabic Language Processing (ALP) Network [44] or the Machine Translation of Arabic (MTRA) project[21].
- **Global Platforms with Arabic Support:** While not exclusive to Arabic, some prominent global platforms offer datasets with Arabic content and user interfaces in Arabic. Here are a few examples:
  - **Hugging Face Datasets:** Explore their extensive collection, filtering by languages and searching for datasets tagged with "Arabic" [18].
  - **Google AI Arabic Datasets:** This initiative by Google AI aims to support research in Arabic NLP. While still under development, it might offer valuable resources in the future.
  - **Kaggle Arabic Datasets:** While Kaggle itself isn't Arabic-specific, some datasets uploaded by the community might be in Arabic. Utilize their search function with keywords like "Arabic" or specific domains like "Arabic sentiment analysis" [20].

**Examples of Arabic-Oriented Datasets on Various Platforms**

The table 2.3 provides examples of Arabic-oriented datasets available on various platforms.For Hugging Face, datasets such as the MADAR Arabic Dialect Corpus, AraVec: Arabic Word Embeddings, Arabic Sentiment Twitter Corpus, and Arabic Wikipedia Text are highlighted. These datasets cover a range of linguistic and text-based applications, including dialect analysis, word embeddings, sentiment analysis, and Wikipedia text processing. In Kaggle, users can find datasets like Arabic Online Shopping Reviews, Arabic Handwritten Digits, Arabic News Text Classification, and Arabic Poetry Dataset. These datasets cater to different domains such as e-commerce, handwriting recognition, news categorization, and poetry analysis, offering diverse opportunities for research and development. Google Dataset Search hosts datasets such as Arabic Speech Recognition, Arabic Named Entity Recognition, Arabic Image Captioning, and Arabic Emotion Recognition. These datasets focus on multimedia and language processing tasks, including speech recognition, named entity recognition, image captioning, and emotion analysis, contributing to advancements in artificial intelligence and natural language processing for Arabic language applications.

| Platform | Examples of Arabic-oriented Datasets |
|---|---|
| Hugging Face | MADAR Arabic Dialect Corpus [22]<br>AraVec: Arabic Word Embeddings [12]<br>Arabic Sentiment Twitter Corpus [9]<br>Arabic Wikipedia Text [11] |
| Kaggle | Arabic Online Shopping Reviews [7]<br>Arabic Handwritten Digits [3]<br>Arabic News Text Classification [6]<br>Arabic Poetry Dataset [8] |
| Google Dataset Search | Arabic Speech Recognition [10]<br>Arabic Named Entity Recognition [5]<br>Arabic Image Captioning [4]<br>Arabic Emotion Recognition [2] |

Table 2.3: Examples of Arabic-oriented datasets in various platforms

### 2.4.2 Challenges in Identifying Dataset Platforms for Arabic Content

Identifying specific dataset platforms dedicated solely to the Arabic world and Arabic content can be challenging due to several reasons:

- **Limited Availability**: There is a scarcity of dedicated dataset platforms focused solely on the Arabic world and Arabic content. Unlike regions with well-established data ecosystems, such as North America or Europe, the Arabic world may have fewer resources dedicated to curating and sharing datasets.
- **Fragmentation**: The Arabic world consists of multiple countries with diverse languages, cultures, and data policies. As a result, dataset platforms may be fragmented across different countries and regions, making it difficult to identify a single platform that caters exclusively to Arabic content.
- **Language Barrier**: Most dataset platforms, especially those with global reach, primarily use English as the primary language for data descriptions, metadata, and documentation. This language barrier may pose challenges for Arabic-speaking users in discovering and accessing relevant datasets.

- **Data Accessibility**: In some cases, datasets specific to the Arabic world may be available but scattered across various sources such as government portals, research institutions, and private organizations. This lack of centralized repositories makes it challenging for users to locate and access Arabic datasets easily.
- **Data Privacy and Regulations**: Data privacy laws and regulations in the Arabic world may vary across countries, affecting the availability and accessibility of datasets. Some datasets, especially those containing sensitive information, may not be publicly accessible or may have restricted access, limiting their availability on public dataset platforms.
- **Awareness and Infrastructure**: There may be a lack of awareness and investment in data infrastructure and open data initiatives in the Arabic world compared to other regions. This can result in fewer dedicated dataset platforms and limited efforts to promote data sharing and collaboration within the region.

Overall, while efforts are being made to promote data sharing and collaboration in the Arabic world, the lack of dedicated dataset platforms and the challenges associated with data accessibility, fragmentation, and language barriers contribute to the difficulty in identifying specific platforms focused solely on Arabic content. Thus, Designing and developing a dedicated dataset platform for AI applications oriented towards Arabic content holds significant promise in overcoming the challenges outlined above. By creating such a platform, researchers, developers, and policymakers in the Arabic-speaking world would gain access to a centralized repository of high-quality datasets tailored to their linguistic and cultural context. This would not only streamline the process of data discovery and accessibility but also foster collaboration and knowledge sharing within the region's AI community. Moreover, a dedicated platform could address issues of data fragmentation by aggregating diverse datasets from various sources and domains, thereby providing a comprehensive resource for AI research and development. Additionally, the platform could implement language localization features to bridge language barriers, making it more accessible to users who primarily speak Arabic. Overall, investing in the creation of a dataset platform specifically focused on Arabic content would be a crucial step towards advancing AI capabilities in the Arabic-speaking world and fostering innovation in various domains, including natural language processing, computer vision, and machine learning.

## 2.5 Summary

This chapter explores the significance of dataset platforms in driving innovation and addressing challenges in the AI domain. It discusses prominent global and regional platforms, highlighting their role in facilitating research and collaboration. Despite challenges in identifying Arabic-focused datasets, efforts are underway to promote data sharing in the Arabic world. The chapter advocates for the development of dedicated dataset platforms for Arabic content to overcome challenges and foster AI advancements. Such platforms can promote collaboration, innovation, and societal impact, contributing to the global AI landscape.

CHAPTER 3

# PLANNING PHASE

## 3.1 Introduction

The planning phase of dataset platform development is a pivotal stage where we lay the groundwork for constructing an effective platform to cater to the evolving demands of data scientists, researchers, developers, and enthusiasts.This chapter aims to delineate our objectives in a precise and comprehensive manner and to delineate the indispensable requirements—both functional and non-functional—as well as user requirements crucial for the platform's success. By establishing a robust foundation during this planning phase, we ensure that our dataset platform not only meets but surpasses the expectations of our diverse user base. Our content strategy will revolve around harmonizing these objectives and requirements with our overarching vision, ensuring that every aspect of the platform development contributes to a seamless and enriching user experience. Let us embark on a journey to explore the intricacies of planning a dataset platform that empowers and inspires its users.

## 3.2 Objectives

Creating a dataset platform for AI purposes with Arabic-oriented content can serve various objectives to cater to the needs of researchers, developers, and organizations interested in AI applications in Arabic language and culture. Here are some objectives for developing such a website:

- **Facilitate Research**: Provide a centralized platform for researchers and academics to access diverse datasets relevant to Arabic language, culture, and demographics. This facilitates research in various AI-related fields such as natural language processing, computer vision, and sentiment analysis.
- **Support AI Development**: Enable developers working on AI projects to access high-quality datasets specifically curated for Arabic language tasks. This can accelerate the development of AI models and applications tailored for Arabic-speaking users.
- **Promote Innovation**: Encourage innovation in AI by offering a wide range of Arabic datasets covering different domains and topics. This allows developers and data scientists to explore new ideas and experiment with cutting-edge AI techniques in an Arabic context.
- **Address Language Specific Challenges**: Address the unique challenges and nuances of the Arabic language in AI development, such as dialectal variations, linguistic features, and cultural contexts. By providing specialized datasets, the platform helps developers overcome these challenges and build more accurate and culturally sensitive AI models.
- **Enhance Language Understanding**: Contribute to the advancement of Arabic language understanding and processing capabilities in AI systems. By curating diverse and comprehensive datasets, the platform supports the training and evaluation of AI models for tasks like machine translation, sentiment analysis, and named entity recognition in Arabic text.
- **Support Multimodal AI**: Include not only textual data but also multimedia datasets (e.g., images, videos, audio) with Arabic content. This enables researchers and developers to work on multimodal AI applications that analyze and generate content in multiple formats, enhancing the richness and versatility of AI solutions for Arabic users.
- **Foster Collaboration**: Facilitate collaboration among researchers, developers, and organizations interested in Arabic-focused AI projects. The platform can include features such as forums, community discussions, and collaboration tools to encourage knowledge sharing, networking, and joint research efforts.

- **Ensure Accessibility and Ethical Use**: Ensure that the datasets hosted on the platform are accessible to a wide range of users while also upholding ethical standards and data privacy principles. This includes providing proper documentation, obtaining necessary permissions for dataset usage, and implementing safeguards to prevent misuse or unethical practices.

## 3.3 Requirements

When it comes to website development, various types of requirements need to be considered to ensure the successful creation and implementation of the website. These requirements can be categorized into several types:

### 3.3.1 Functional Requirements

Designing and developing a dataset platform for AI purposes with an Arabic-oriented content involves considering several functional requirements to ensure the platform meets the needs of users and effectively serves its intended purpose. Here are some functional requirements to consider:

- **User Authentication and Authorization**:
  - Allow users to register accounts and log in securely.
  - Implement role-based access control to manage user permissions (e.g., administrators, contributors, viewers).
- **Dataset Management**:
  - Enable users to upload, organize, and manage datasets.
  - Support various data formats commonly used in AI applications (e.g., CSV, JSON, image files).
  - Provide metadata fields for each dataset (e.g., title, description, tags, licensing information).
- **Search and Discovery**:
  - Implement a robust search functionality to allow users to discover datasets based on keywords, categories, or tags.

  - Include advanced search filters to narrow down results (e.g., by data type, date range, language).
  - Support both Arabic and English language search queries.

- **Dataset Preview and Exploration**:
  - Allow users to preview dataset contents, including sample data and metadata, before downloading.
  - Provide visualization tools to explore data distributions, correlations, and patterns.
  - Support interactive exploration features such as filtering, sorting, and aggregation.

- **Data Quality Assurance**:
  - Implement tools for data validation and quality assessment to ensure datasets are accurate, consistent, and reliable.
  - Allow users to report issues or errors with datasets and collaborate on data improvement efforts.

- **Dataset Versioning and History**:
  - Enable version control to track changes made to datasets over time.
  - Maintain a revision history to allow users to revert to previous versions if needed.
  - Provide audit logs to track dataset modifications and user actions.

- **Collaboration and Community Engagement**:
  - Facilitate collaboration among users by allowing them to share datasets, provide feedback, and contribute annotations or labels.
  - Incorporate social features such as comments, ratings, and user reviews to foster community engagement.

- **API and Integration**:
  - Provide a RESTful API to allow programmatic access to dataset metadata and contents.

  - Support integration with popular AI frameworks, libraries, and tools (e.g., TensorFlow, PyTorch) for seamless data ingestion and preprocessing.

- **Performance and Scalability**:
  - Ensure the platform can handle large volumes of data and concurrent user requests.
  - Implement caching mechanisms and optimize database queries for improved performance.
  - Design the platform to be horizontally scalable to accommodate future growth.

- **Security and Compliance**:
  - Implement security best practices to protect user data and prevent unauthorized access.
  - Encrypt data transmissions and enforce secure authentication mechanisms.
  - Ensure compliance with data privacy regulations (e.g., GDPR) and obtain necessary consent for data usage.

### 3.3.2 Non Functional Requirements

Non-functional requirements (NFRs) for a website define how the site should behave and operate, rather than what specific features it should have. These indirectly impact the user experience and overall success of the website. Here are some common types of website non-functional requirements.

| Category | Description |
|---|---|
| Performance | Fast load times and responsiveness for data access and visualization |
| Scalability | Ability to handle large and growing datasets efficiently |
| Availability | High uptime to ensure continuous availability of data for AI applications |
| Security | Robust security measures to protect sensitive data from unauthorized access |
| Usability | Intuitive interface for data exploration, analysis, and model training |
| Maintainability | Easy to maintain and update to accommodate new data sources and functionalities |
| Interoperability | Ability to integrate with different AI frameworks and tools seamlessly |

Table 3.1: Non functional requirement of dataset platform

### 3.3.3 User Requirements

User requirements for a dataset platform for AI applications focus on how users will interact with the platform to achieve their goals. Here are some key areas to consider:

- Search and Filter datasets (keywords, type, domain).
- Preview and Download data (various formats).
- Upload datasets with versioning.
- Data annotation and labeling (user-friendly).
- Collaboration features (shared data, reviews, discussions)
- Intuitive interface for diverse users.
- User roles and permissions management

Our target audience encompasses a diverse spectrum of individuals deeply engaged or aspiring to delve into the realm of artificial intelligence. For seasoned professionals such as data scientists and machine learning engineers, we offer advanced tools and insights to sharpen their expertise and navigate the cutting edge of AI innovation. Data analysts and researchers find our resources invaluable for discovering new methodologies and refining their analytical approaches. Developers dedicated to crafting AI-driven applications rely on our platform for robust frameworks and practical guidance to realize their visions. Moreover, we warmly welcome students and beginners eager to embark on their AI journey, providing a nurturing environment rich with foundational knowledge and hands-on learning opportunities to ignite their passion for the field.

## 3.4 Business Model

A business model outlines how a startup creates, delivers, and captures value. It encompasses the organization's core strategy, revenue streams, cost structure, and target market. A well-defined business model clarifies how a startup generates profit, helps identify potential risks and opportunities, guides resource allocation, and facilitates strategic decision-making. Moreover, it provides a framework for innovation and adaptation in response to changing market dynamics, ensuring the long-term viability and success of the business. The business model consists of nine building blocks categorized into five main sections: Customer Segments, Value Propositions, Channels, Customer

Relationships, and Revenue Streams. The right side of the canvas focuses on value creation and delivery, while the left side focuses on the customer segments. Here's a breakdown of the specific elements within the business model for the Arabic-oriented data set platform as presented in the following figure 3.1.

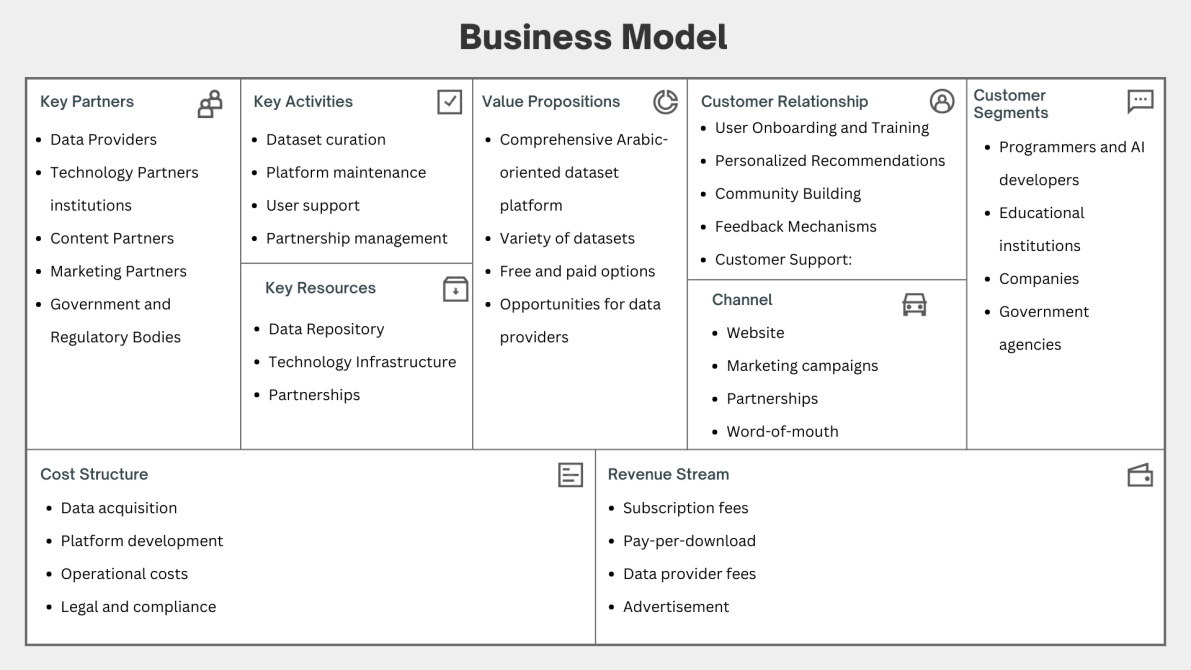


Figure 3.1: The business model

## 3.5 Summary

This chapter outlines the objectives, requirements, and Business model for developing a dataset platform tailored for AI applications focusing on Arabic language and culture. Objectives include facilitating research, supporting AI development, promoting innovation, addressing language-specific challenges, and ensuring ethical data use. Functional requirements encompass user authentication, dataset management, search and discovery, collaboration features, performance, security, and more. Non-functional requirements cover aspects like performance, scalability, security, usability, and interoperability. User requirements focus on interactions such as search, data preview, upload, annotation, collaboration, and user management.

CHAPTER 4

# CONCEPTION PHASE

## 4.1 Introduction

This chapter provide a holistic view of the design and conceptualization processes essential for bringing ideas to fruition in the digital realm. It delves into the pivotal Conception Phase, laying the groundwork for the development journey ahead. This chapter navigates through the intricacies of both Front-end and Back-end conceptualization. In Section 4.2, Front-end Conceptualization explores the essential elements such as Sitemaps, Wireframes, and Initial Prototypes, which serve as the blueprint for user interaction and interface design. Meanwhile, Section 4.3 delves into Back-end Conception, meticulously detailing Platform Architecture and Database Description, crucial components shaping the foundation of the project's technical infrastructure.

## 4.2 Front-end Conceptualization

Front end conceptualization refers to the process of planning and designing the user interface and user experience of a digital product or platform. It aims to ensure that the digital product is user-friendly, visually appealing, and effectively communicates its purpose and features to the users. It involves understanding the needs and expectations of the target users and translating those insights into a coherent and intuitive interface design. This process typically includes creating sitemaps to organize the structure of the platform, developing wireframes to outline the layout and functionality of each screen,

and building prototypes to visualize and test the user flow and interaction design.

### 4.2.1 Sitemap

A site map is a hierarchical representation of the structure of a website, organized in a tree-like format. It outlines the relationships between different pages and content elements within the site, showing how they are interconnected and arranged. This hierarchical structure helps users and developers understand the organization and navigation of the website, making it easier to find and access specific content. Additionally, site map trees can serve as a blueprint for website design and development.

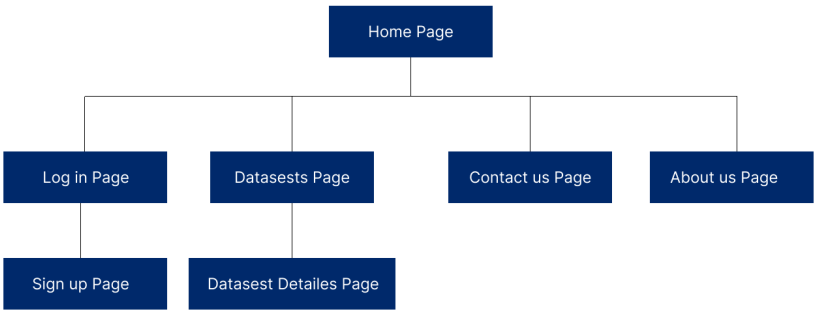


Figure 4.1: Site map for Platform Visitor

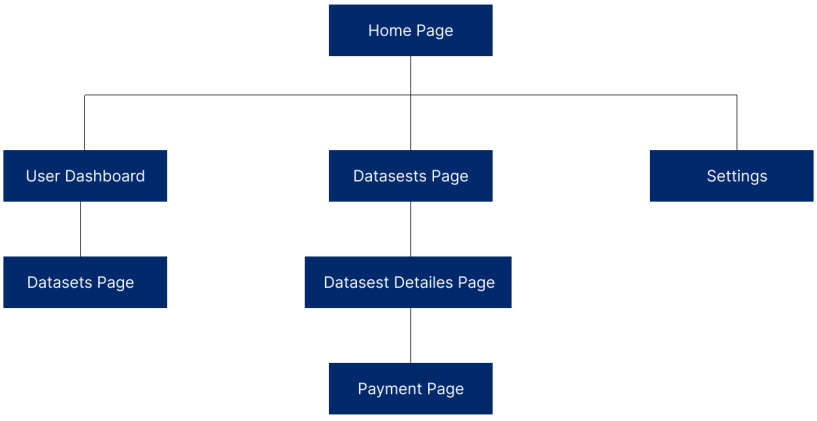


Figure 4.2: Site map for User after Sign up

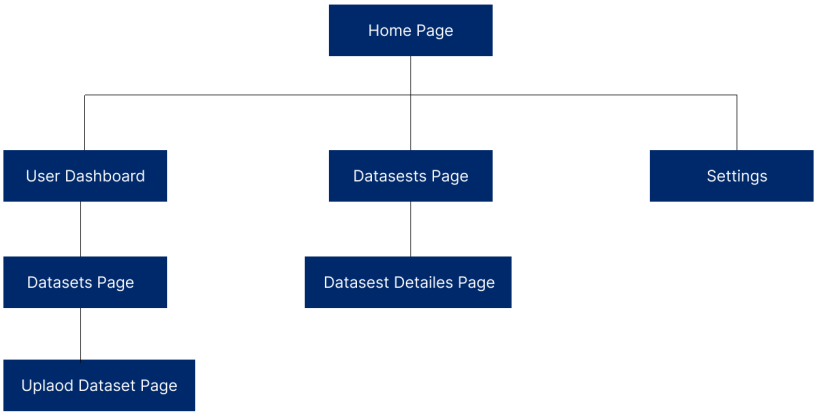


Figure 4.3: Site map for Provider after Log in

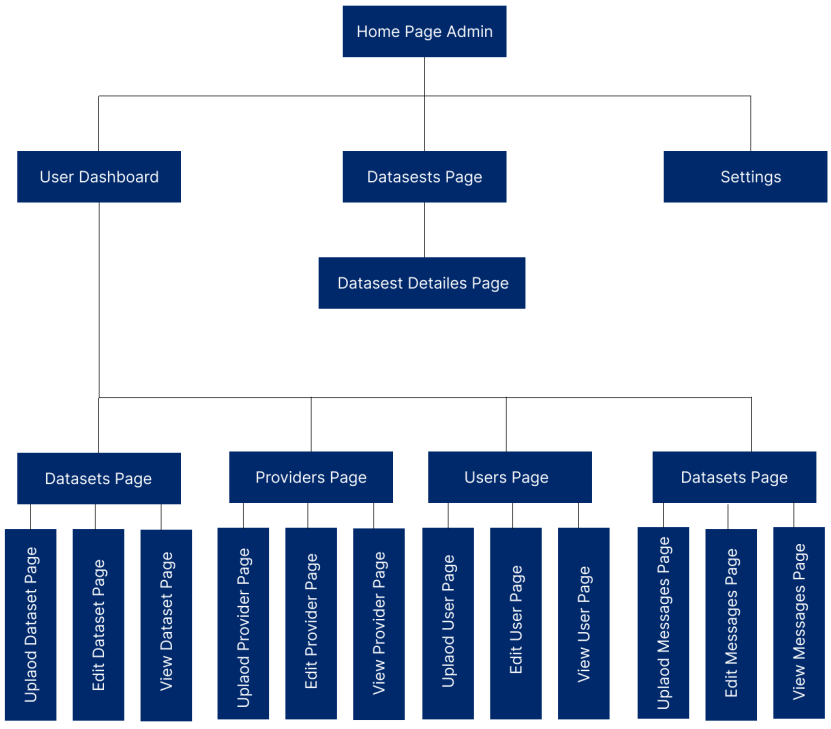


Figure 4.4: Site map for Admin after Log in

### 4.2.2 Wireframes

A wireframe is a visual representation or blueprint of a web page or application interface that outlines the basic layout and structure without including detailed design elements or content. It typically consists of simple shapes, lines, and placeholders to represent different elements such as text, images, buttons, and navigation menus. Wireframes focus on defining the overall arrangement of elements on the screen, including their relative placement, sizing, and functionality, while omitting specific visual details such as colors, fonts, and images. They serve as a valuable tool for conceptualizing and communicating design ideas, facilitating collaboration among designers, developers, and stakeholders, and iterating on the user interface before finalizing the visual design.

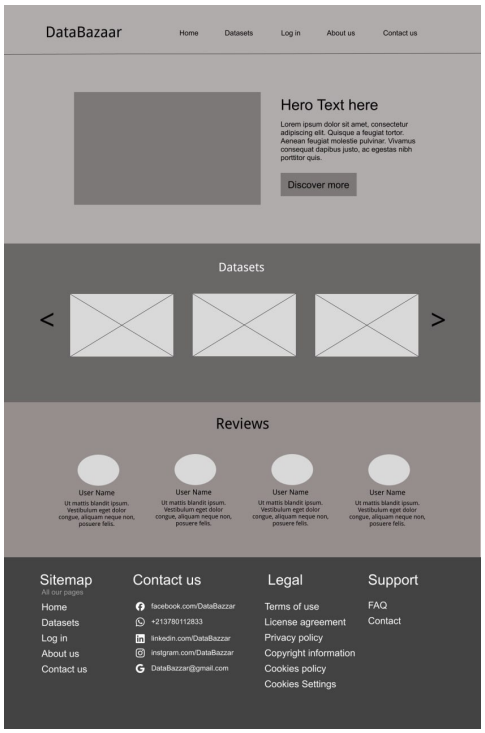


Figure 4.5: Wire frame of Homepage

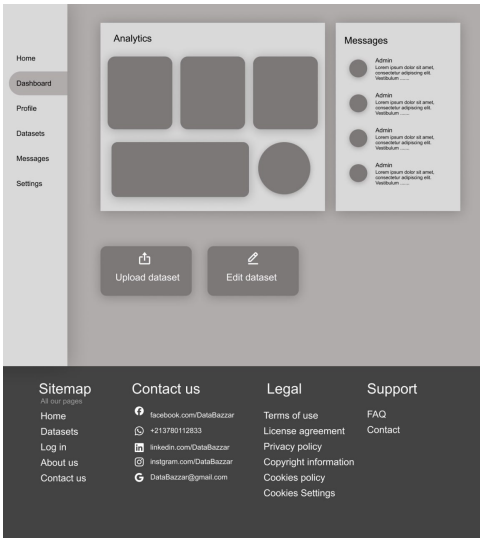


Figure 4.6: Wire frame of dashboard

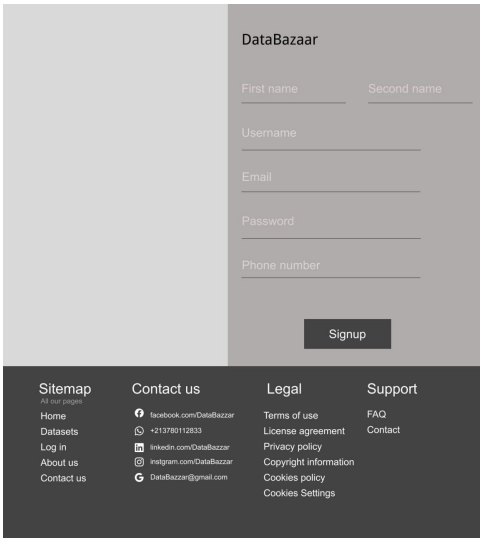


Figure 4.7: Wire frame of sign up form

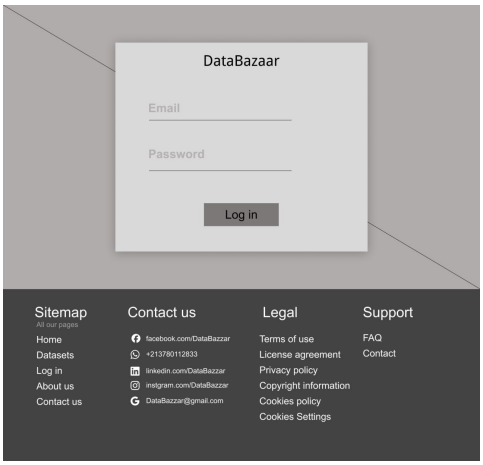


Figure 4.8: Wire frame of login Form


Figure 4.9: Wire frame of dataset description page

### 4.2.3 Initial Prototypes

Prototypes of websites are preliminary versions or mockups created during the initial stages of web development to visualize the layout, design, and functionality of a website before its full implementation. These prototypes serve as a blueprint or a roadmap for the final product, allowing designers and developers to experiment with different ideas, gather feedback, and refine the user experience. Prototypes can range from simple wireframes outlining basic page structures to high-fidelity mockups with detailed design elements and interactive features. By creating prototypes, stakeholders can better understand the project requirements, make informed decisions, and identify potential issues early in the development process, ultimately leading to the creation of a more polished and user-friendly website .

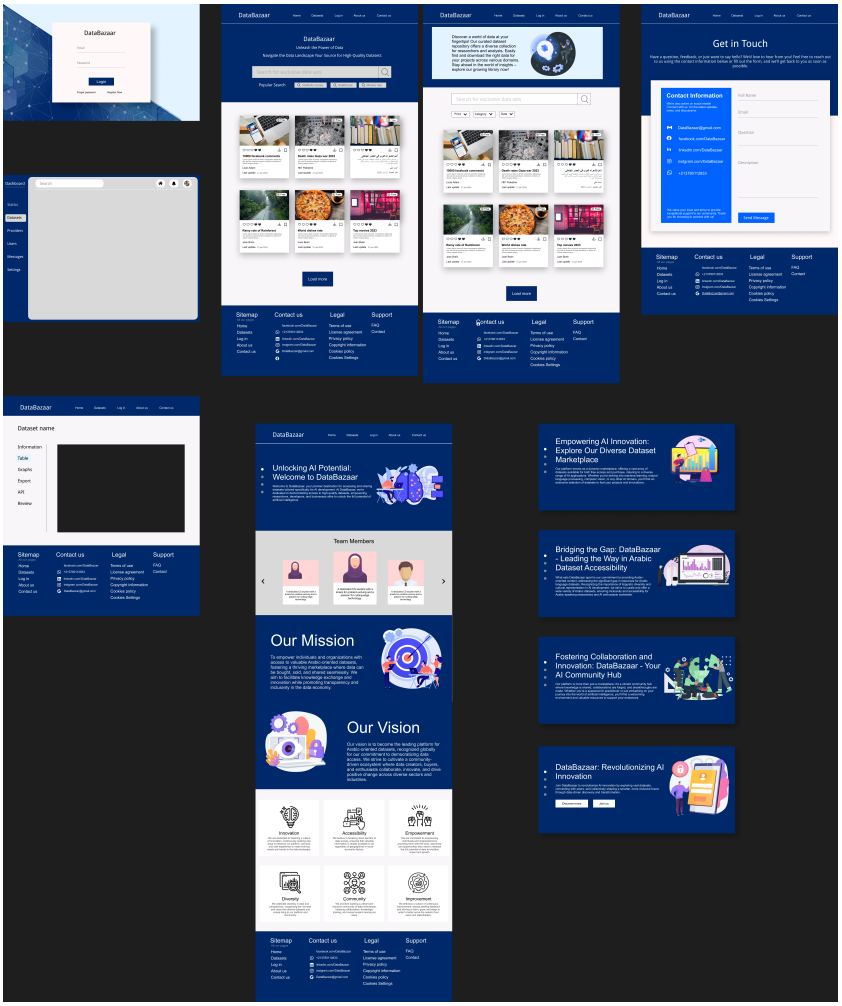


Figure 4.10: Initial Prototype of the platform created in Figma framework

## 4.3 Back-end Conception

The conception of the backend for the DataBaazar platform is a critical phase that ensures the robust functionality and seamless integration of various components.

### 4.3.1 Platform Architecture

Unified Modeling Language (UML) is a standardized visual language used to model and document the design of software systems. It provides a set of diagrams and symbols to represent the structural and behavioral aspects of a system. UML is crucial for describing the architecture of a platform as it facilitates clear communication among developers, designers, and stakeholders. By using UML, complex system components and interactions can be visually mapped out, ensuring a comprehensive understanding of the system's functionality and structure. This helps in identifying potential issues early, improving design quality, and streamlining the development process. Utilizing Unified Modeling Language (UML) diagrams, such as use case, class, and sequence diagrams, provides a structured approach to visualize and design the system architecture. Together, these UML diagrams add significant value by providing a comprehensive blueprint of the backend architecture, facilitating clear communication among stakeholders, and guiding the development process to ensure a scalable, efficient, and maintainable platform.

#### 4.3.1.1 Use case diagrams

Use case diagrams play a pivotal role in identifying and representing the interactions between users (actors) and the system, highlighting the primary functionalities and user requirements. This section presents the use case diagrams for the three main actors in the platform: the provider, the user, and the admin.

- **Admin use cases**

The admin is a pivotal actor in the platform, responsible for overseeing and managing the overall system operations. The admin's role includes tasks such as managing user accounts, datasets and moderating content uploaded by providers.

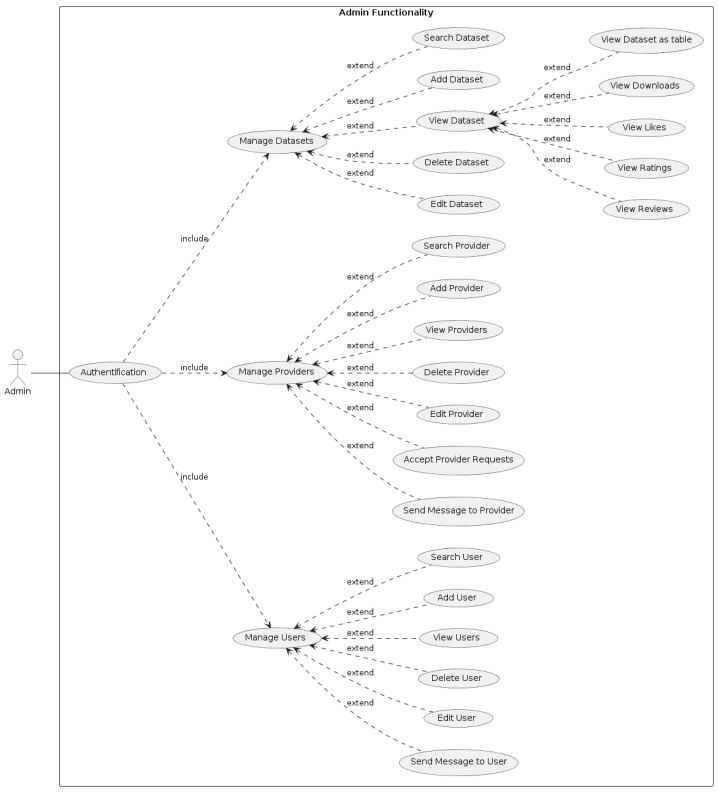


Figure 4.11: Admin use case diagram

- **Provider use cases**

The provider is an essential actor on the platform, responsible for supplying and maintaining datasets. Providers contribute to the platform by uploading datasets, ensuring they are accurate, comprehensive, and well-documented. They play a crucial role in enriching the platform's data repository, making high-quality data available for users and AI developers.

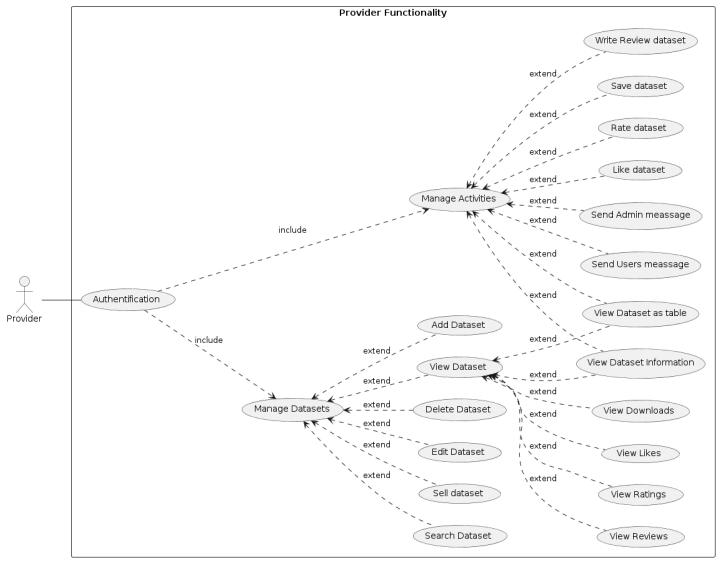


Figure 4.12: Provider use case diagram

- **User use cases**

The user is a key actor on the platform, primarily engaged in searching for, accessing, and utilizing the datasets provided. Users interact with the platform to find relevant data for various purposes, such as research, AI model training, or data analysis. They rely on the quality and accuracy of the datasets and can provide feedback or rate the datasets to help maintain high standards. Users may also contribute by sharing insights or improvements related to the data, fostering a collaborative environment that enhances the overall utility and reliability of the platform's resources.

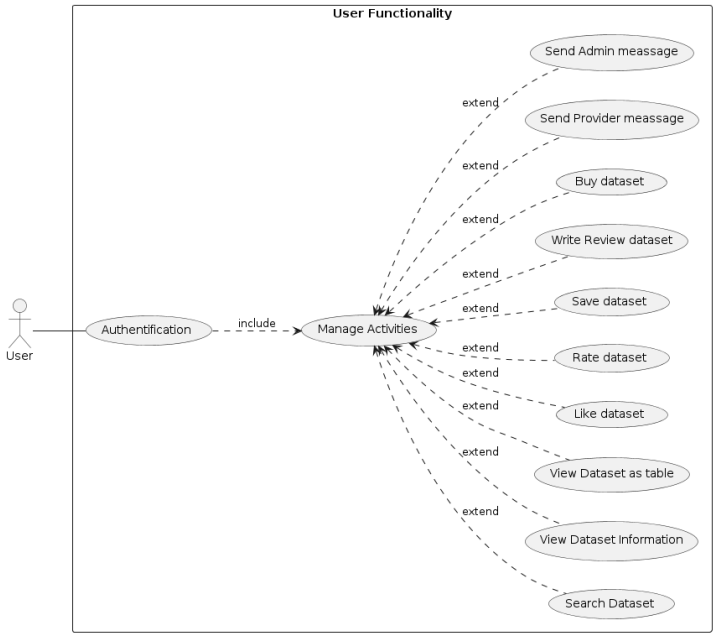


Figure 4.13: User use case diagram

#### 4.3.1.2 Class diagrams

Class diagrams offer a detailed depiction of the system's static structure, illustrating the classes, attributes, methods, and the relationships between them, which is essential for defining the database schema and ensuring data integrity.

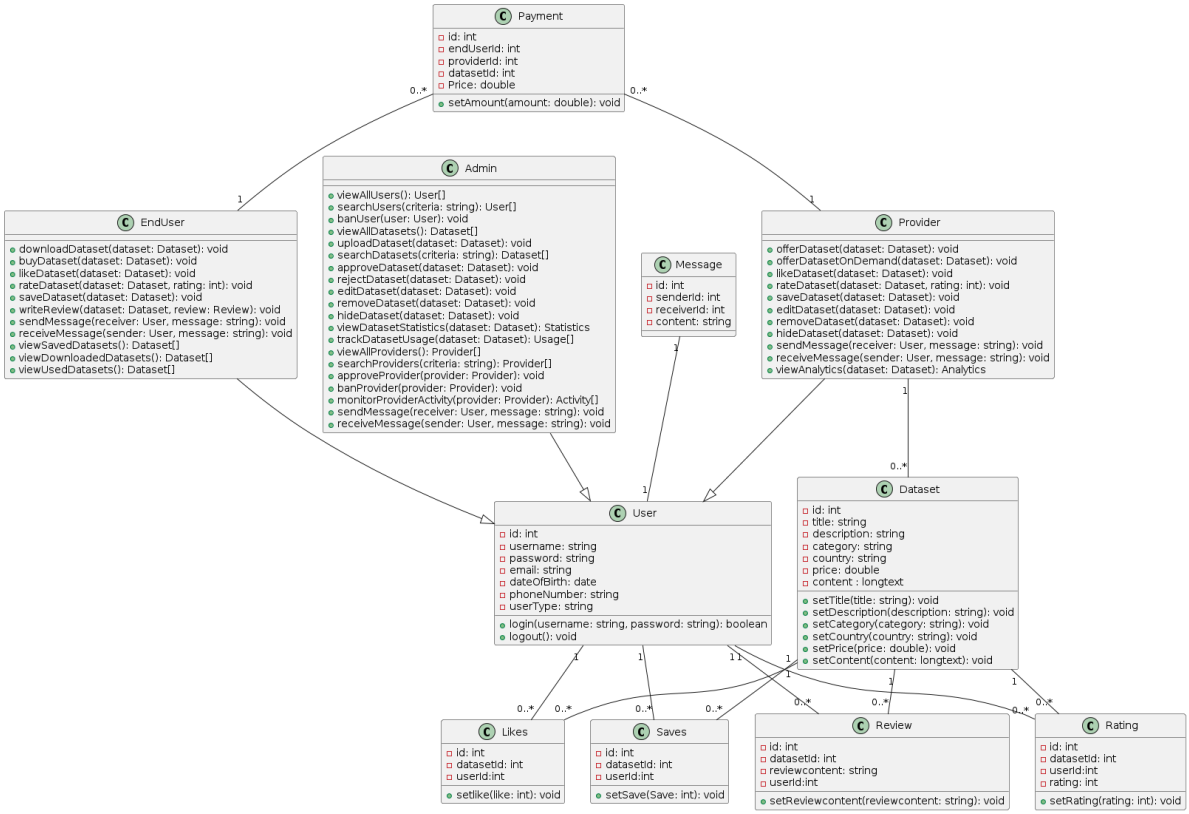


Figure 4.14: Class Diagram

#### 4.3.1.3 Sequence diagrams

Sequence diagrams model the dynamic behavior of the system by showing the sequence of messages exchanged between objects to carry out a specific functionality, thereby elucidating the workflow and interaction logic.

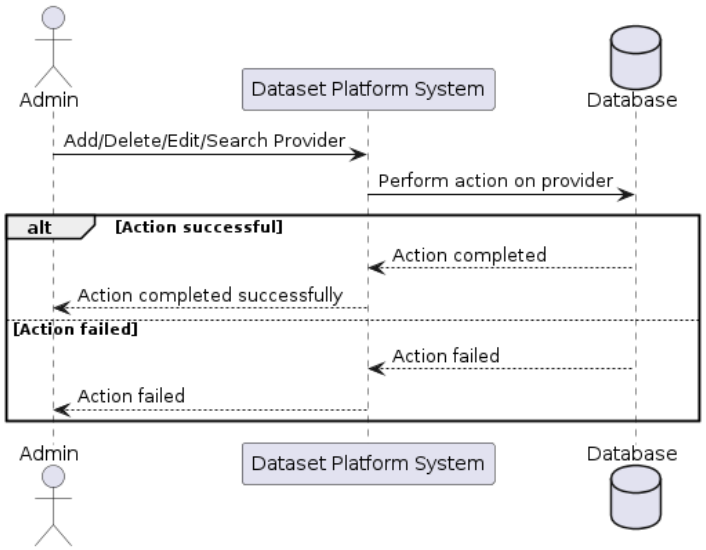


Figure 4.15: Sequence Diagram of Admin Provider Interaction

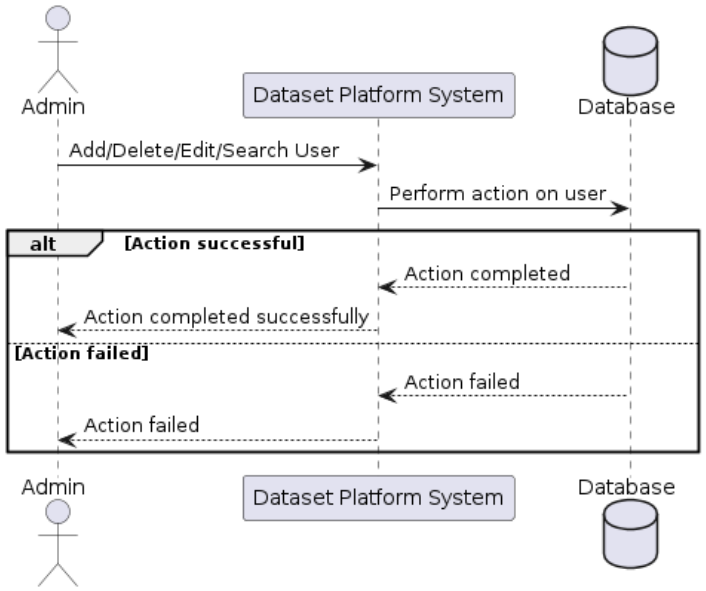


Figure 4.16: Sequence Diagram of Admin User Interaction

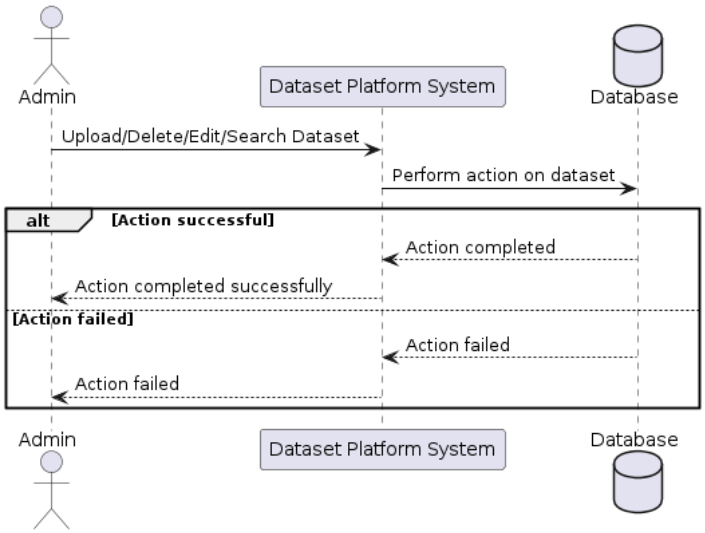


Figure 4.17: Sequence Diagram of Admin Dataset

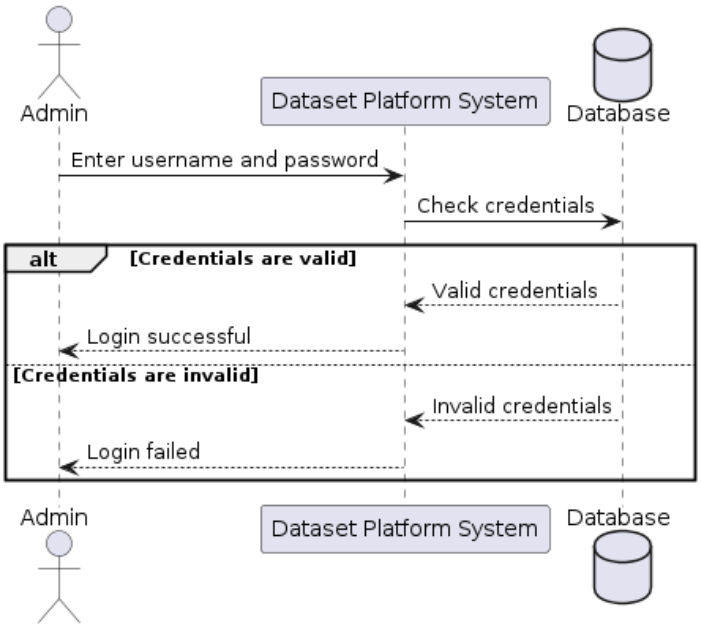


Figure 4.18: Sequence Diagram of Admin login

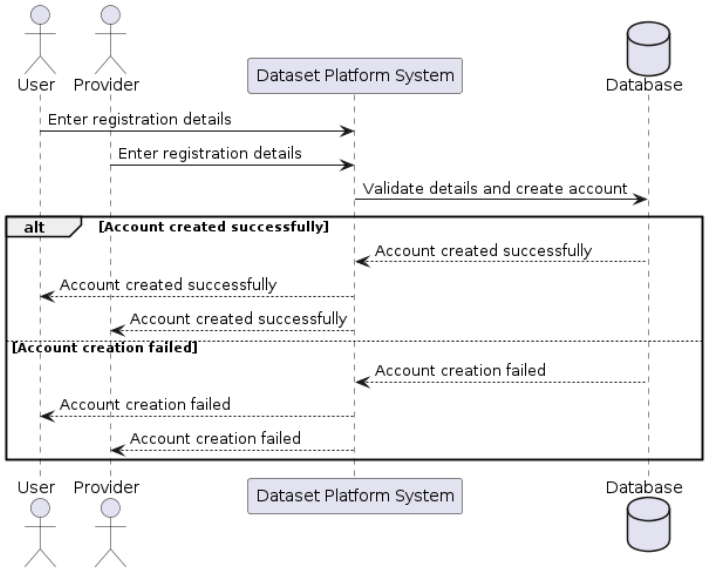


Figure 4.19: Sequence Diagram of Provider/user Sign up

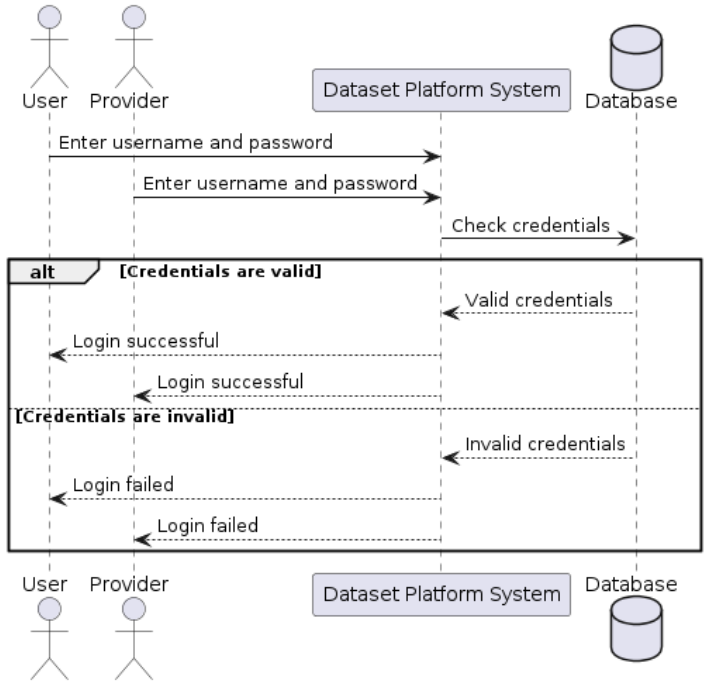


Figure 4.20: Sequence Diagram of Provider/ User Signup Provider-user

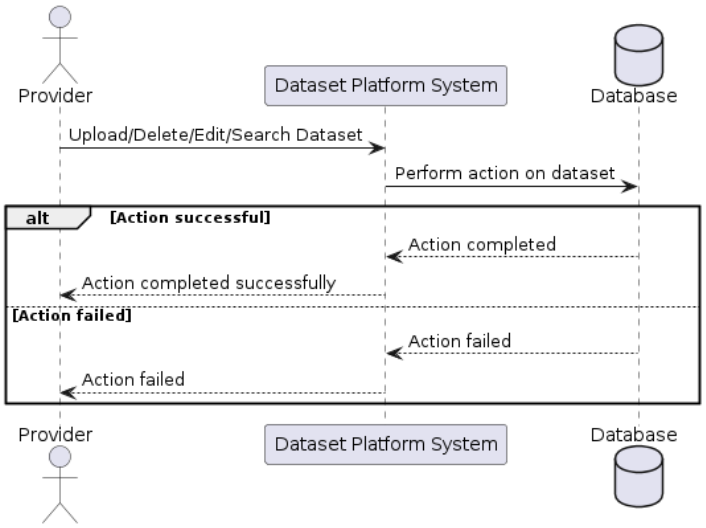


Figure 4.21: Sequence Diagram of Provider Dataset Interaction as provider

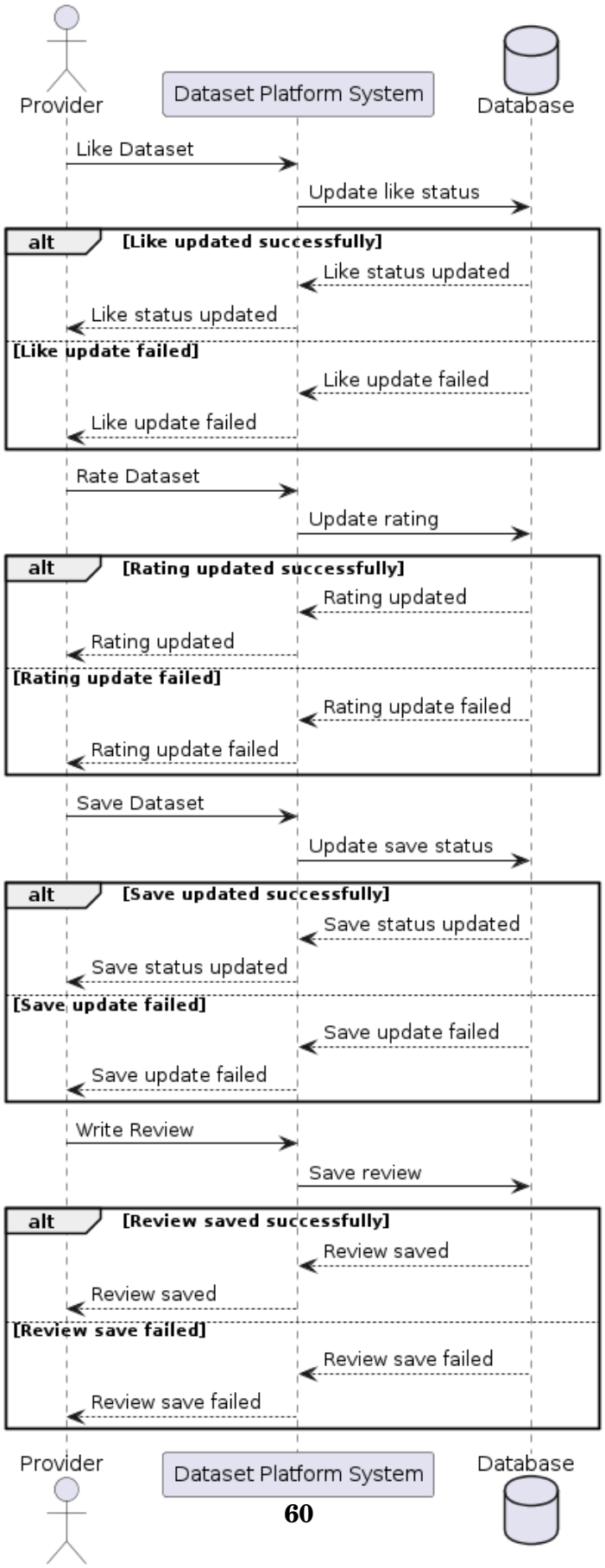




Figure 4.22: Sequence Diagram of Provider Dataset Interaction as End-User

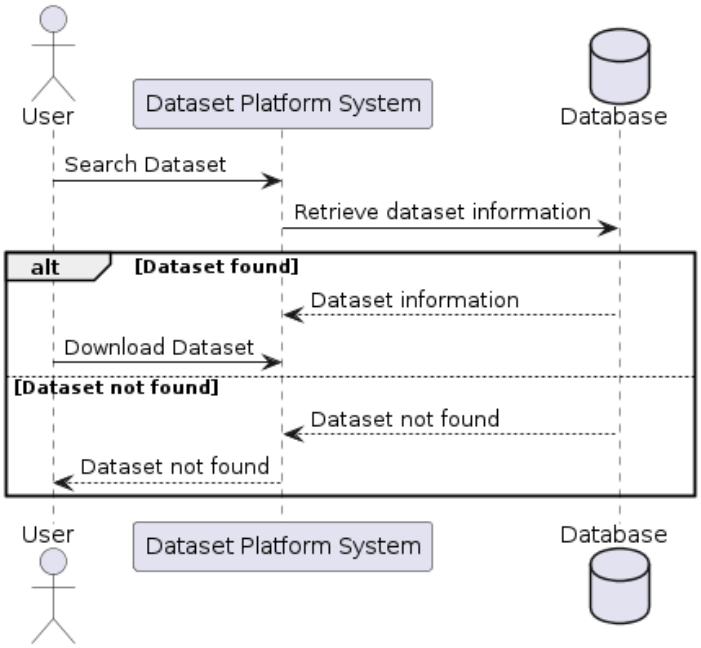


Figure 4.23: Sequence Diagram of User search

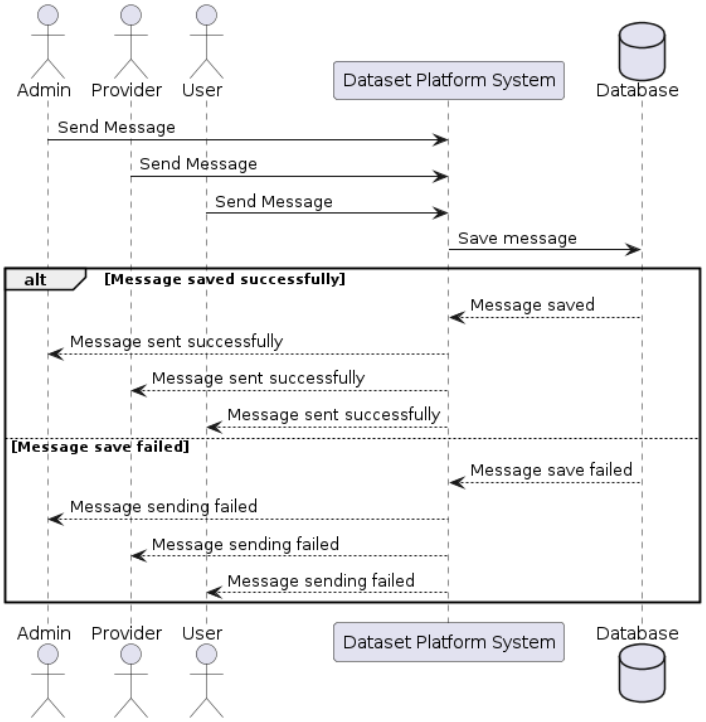


Figure 4.24: Sequence Diagram of Sending message Use Case

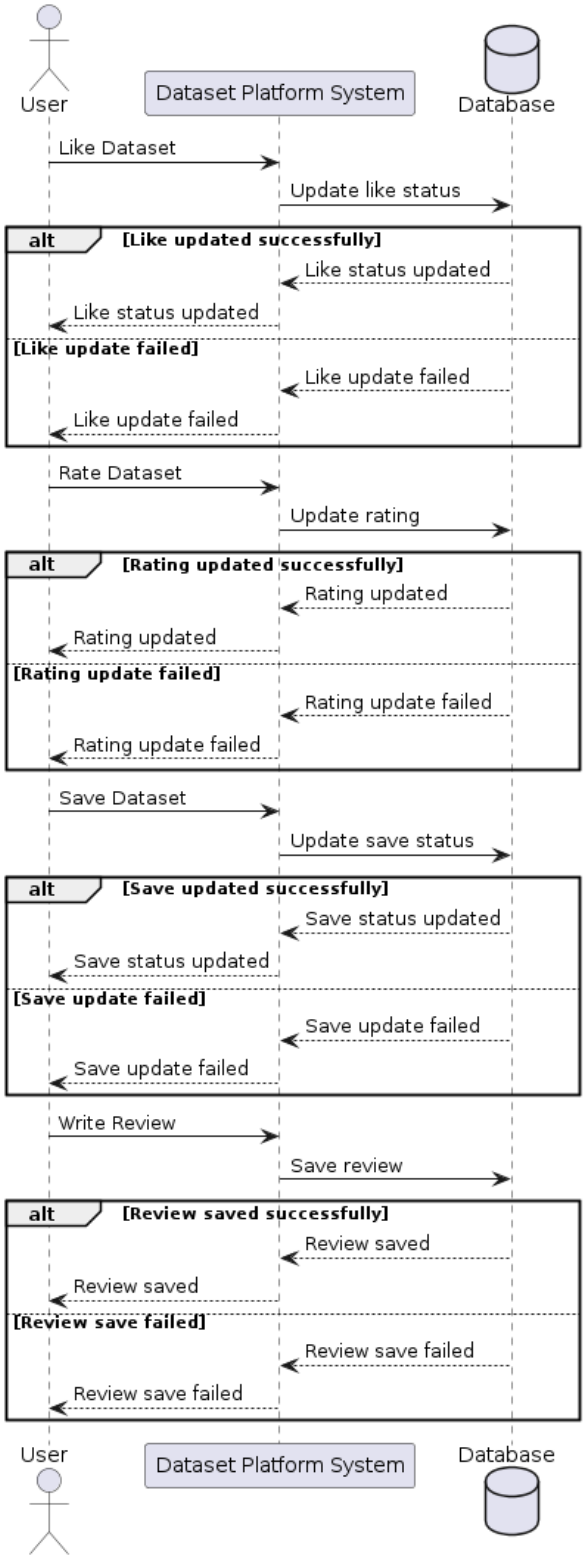


Figure 4.25: Sequence Diagram of User Dataset Interaction

### 4.3.2 Database Description

Our database comprises several interconnected tables designed to support a dataset platform. The "User Table" stores essential user information such as usernames, passwords, emails, and user types. Each user can interact with datasets, which are stored in the "Dataset Table" along with details like title, description, category, price, and content. Reviews and ratings for datasets are captured in separate "Review" and "Rating" tables, linked to users and datasets by unique identifiers. Additionally, users can express their preferences through the "Likes" and "Saves" tables, indicating interest in specific datasets. Communication between users is facilitated by the "Message Table," recording sender and receiver information alongside message content. Finally, transactions are logged in the "Payment Table," tracking payments made by end users to dataset providers. This database structure forms the backbone of the platform, enabling efficient management, analysis, and interaction with diverse datasets while maintaining user engagement and facilitating transactions.

- **User Table**

| id | username | password | email | dateOfBirth | phoneNumber | userType | country |
|---|---|---|---|---|---|---|---|

- **Dataset Table**

| id | title | description | country | category | price | content | providerId |
|---|---|---|---|---|---|---|---|

- **Review Table**

| id | datasetId | reviewcontent | userId |
|---|---|---|---|

- **Rating Table**

| id | datasetId | userId | rating |
|---|---|---|---|

- **Likes Table**

| id | datasetId | userId |
|---|---|---|

- **Saves Table**

| id | datasetId | userId |
|---|---|---|

- **Message Table**

| id | senderId | receiverId | content |
|---|---|---|---|

- **Payment Table**

| id | endUserId | providerId | datasetId | Price |
|---|---|---|---|---|

## 4.4 Summary

In this chapter, the focus is on laying the groundwork for the project's development journey. It guides through both the Front-end and Back-end conceptualization phases, which are essential for realizing ideas. Front-end Conceptualization, as detailed in Section 4.2, encompasses elements like Sitemaps, Wireframes, and Initial Prototypes, providing blueprints for user interaction and interface design. Meanwhile, Section 4.3 delves into Back-end Conception, meticulously detailing Platform Architecture and Database Description, essential components shaping the project's technical infrastructure. With detailed use case diagrams, class diagrams, sequence diagrams, and a comprehensive database structure, this chapter ensures a holistic understanding of the project's foundational aspects, setting the stage for successful development and implementation.

CHAPTER 5

# IMPLEMENTATION AND DEVELOPMENT

## 5.1 Introduction

The development and implementation phase is crucial in creating a robust dataset platform, as it transforms conceptual designs into functional, user-friendly tools. This phase ensures that the platform's infrastructure is sound, secure, and scalable, addressing both front-end and back-end requirements. Through meticulous implementation, the platform can effectively curate and manage a variety of datasets, providing seamless access to users. Additionally, this phase focuses on optimizing user experience and interface, making the platform intuitive and efficient for diverse user groups, including programmers, AI developers, educational institutions, companies, and government agencies. By meticulously planning and executing each development step, the platform can offer reliable services, accommodate growth, and adapt to evolving user needs, ultimately ensuring its success and sustainability in the competitive data marketplace

## 5.2 Implementation Tools

Implementation in website development refers to the phase where the planned features, designs, and functionalities are put into action. It involves coding, designing, integrating various elements, and building the website according to the specifications outlined during the planning stage. Implementation encompasses translating the conceptual aspects

of the website into tangible digital components, ensuring that the website functions smoothly and meets the intended objectives and user requirements. In this section, we will delineate the various tools crucial for website development, categorizing them into graphic design tools, programming tools, and additional tools such as APIs and libraries. Graphic design tools encompass software like Figma, Plant UML, used for creating visually appealing layouts, images, and graphical elements. Programming tools refer to software environments like integrated development environments (IDEs) such as Visual Studio Code and XAMPP, facilitating coding tasks in languages like HTML, CSS, JavaScript, and others. Additionally, we will explore additional tools like Application Programming Interfaces (APIs) and libraries which provide pre-built functionalities and data integrations, streamlining development processes and enhancing the website's capabilities.

### 5.2.1 Graphic design Tools

In this section, we introduce three essential graphical tools for website development: Figma, PlantUML, and Draw.io, each serving distinct purposes in the design and planning phase.

- **Figma** [42]: Figma is a versatile design tool renowned for its collaborative features and user-friendly interface. It excels in creating wire frames, prototypes and site maps, providing a platform for designers and developers to brainstorm and iterate on website layouts and structures collaboratively. With Figma, teams can create interactive prototypes, define user flows, and visualize the overall website architecture, fostering efficient communication and alignment throughout the design process.

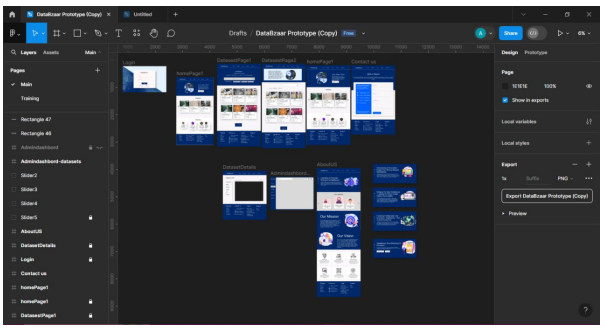

Figure 5.1: Graphical Interface of Figma platform

- **PlantUML** [50]: PlantUML is a textual-based tool used primarily for creating Unified Modeling Language (UML) diagrams. With its simple syntax, developers can generate various UML diagrams, including class diagrams, sequence diagrams, and activity diagrams, to depict the system architecture and functionalities of the website. PlantUML's lightweight nature and integration capabilities make it a preferred choice for documenting and visualizing complex system designs in a concise and standardized manner.

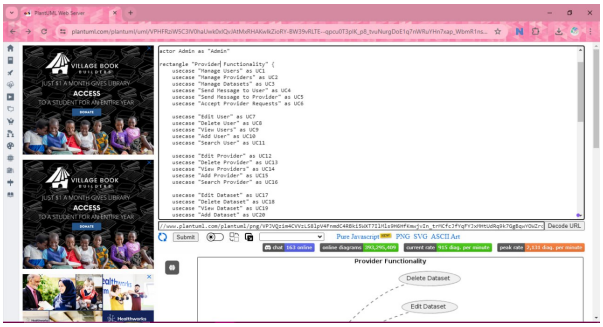


Figure 5.2: Graphical Interface of PlantUML platform

- **Draw.io** [37]: Draw.io is a web-based diagramming tool ideal for creating illustrations and diagrams for website development projects. It offers a wide range of shapes, icons, and templates, allowing users to design custom illustrations, flowcharts, and diagrams to convey concepts, processes, and data structures effectively. Draw.io's intuitive interface and extensive library of elements make it suitable for generating visual assets and diagrams tailored to the specific needs of website design and development projects.

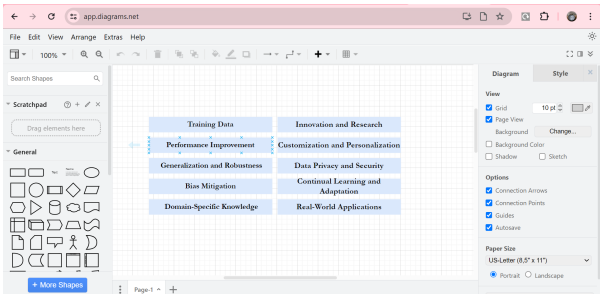


Figure 5.3: Graphical Interface of Drawio platform

### 5.2.2 Programming Tools

In this section, we present various programming tools commonly used in website development as illustrated in Table 5.1. Web development relies on a combination of programming languages to bring digital creations to life. HTML forms the backbone, providing structure and content organization. CSS adds style and visual appeal, shaping the layout and design. JavaScript injects interactivity, enabling dynamic features and responsive behavior. PHP operates behind the scenes, powering server-side functionality and database interactions. Together, these languages synergize to craft immersive, user-friendly web experiences, from sleek interfaces to complex web applications, driving the modern digital landscape forward 5.4.


Figure 5.4: Programming languages

Programming frameworks are pre-established structures, libraries, and tools that provide developers with a foundation for building software applications. Visual Studio, a comprehensive integrated development environment (IDE), supports various languages and frameworks, enabling efficient development of dynamic web applications. XAMPP simplifies local web development by bundling Apache, MySQL, PHP, and Perl into a pre-configured environment. Laravel, a PHP web application framework, emphasizes developer productivity and code maintainability, offering built-in functionalities such as routing and authentication. Together, these frameworks provide developers with the tools needed to streamline web development processes and build scalable, feature-rich applications 5.5.


Figure 5.5: Programming frameworks

### 5.2.3 Additional Tools

- **Swiper.js Library**

Swiper.js is a modern, free, and highly customizable JavaScript library for creating responsive and touch-enabled sliders, carousels, and galleries. It is widely used in web development for its smooth animations, extensive configuration options, and compatibility with both desktop and mobile devices. Swiper.js supports vertical and horizontal sliding, multiple slide layouts, lazy loading, parallax effects, and more, making it a popular choice for interactive and visually appealing web interfaces.

- **SheetJS Library**

SheetJS is a versatile JavaScript library designed for parsing, reading, and writing various spreadsheet formats, including Excel (XLS, XLSX), CSV, and more. It enables developers to handle spreadsheet operations directly in web browsers or Node.js environments, making it a powerful tool for both client-side and server-side applications.

- **Bootstrap framework**

A popular open-source front-end framework used for developing responsive and mobile-first websites. It includes HTML, CSS, and JavaScript components for creating a variety of web elements, such as navigation bars, buttons, forms, and modals.It Widely used to streamline web development by providing pre-designed components and a grid system that ensures a consistent and adaptable layout across different devices.

| Programming Language | Definition |
|---|---|
| HTML (HyperText Markup Language) | The building blocks of web pages, defining structure and content. |
| CSS (Cascading Style Sheets) | Defines the presentation of HTML content, controlling layout, colors, fonts, etc. |
| JavaScript (JS) | A versatile language used for web page behavior, interactivity, and client-side scripting. |
| PHP (Hypertext Preprocessor) | A server-side scripting language for creating dynamic web content. |
| **Programming Framework** | **Definition** |
| Visual Studio | An Integrated Development Environment (IDE) from Microsoft that supports various programming languages and offers tools for code editing, debugging, and project management. |
| Laravel | A PHP framework for building web applications, offering features like routing, authentication, and database access. |
| XAMPP | Not a framework, but a free Apache distribution containing MySQL, PHP, and Perl for running a web server environment on your computer. |

Table 5.1: List of programming tools

## 5.3 DataBaazar development

### 5.3.1 Platform pages

In this section, we will take a detailed tour through the various pages of our platform DataBazaar.

- **Home page**

**A landing page** is a standalone web page designed with a specific purpose in mind, often to encourage visitors to take a particular action, such as signing up for a service, making a purchase, or providing contact information. It typically has a singular focus, presenting concise information and a clear call-to-action (CTA) to guide visitors towards the desired outcome. In the context of our platform, it's important to note that our homepage can be considered as a landing page composed of four distinct sections. Each section serves as a

landing area, providing targeted information and CTAs to engage visitors and encourage them to explore further or take specific actions. By segmenting the homepage into these sections, we can effectively cater to different user interests or goals, optimizing the chances of conversion and enhancing the overall user experience.

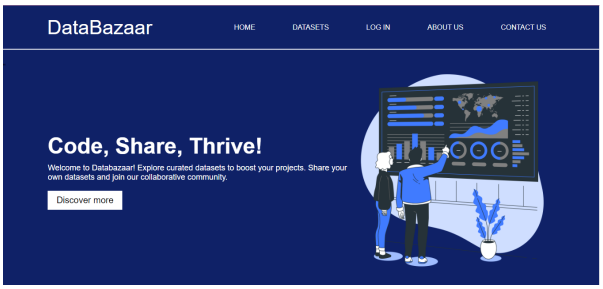


Figure 5.6: Area 1 in home page: Navbar and welcome section

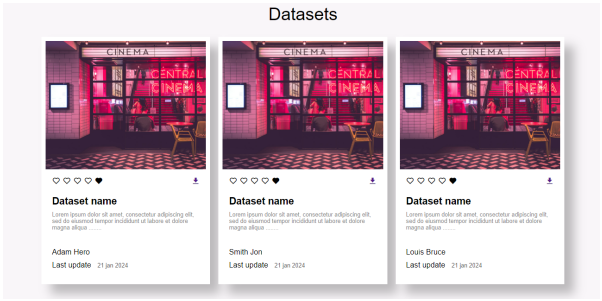


Figure 5.7: Area 2 in home page: Dataset Gallery

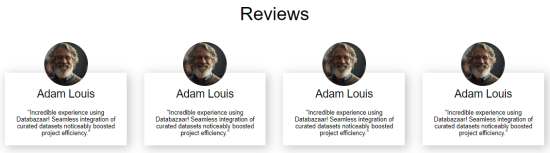


Figure 5.8: Area 3 in home page: Reviews section

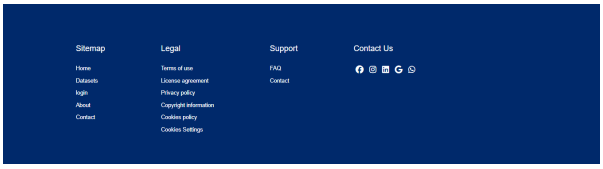


Figure 5.9: Area 4 in home page: Footer section

- **Admin Dashboard**

In the context of a dataset platform, a dashboard is a user interface that provides visualizations, analytics, and insights derived from available datasets. It serves as a control panel or command center, offering a comprehensive overview of various aspects, offering features such as visualization tools, filtering options, and customizable dashboards. Users leverage the dashboard to gain insights, track metrics, and make data-driven decisions effectively, while also facilitating data management, collaboration, and sharing within the platform.

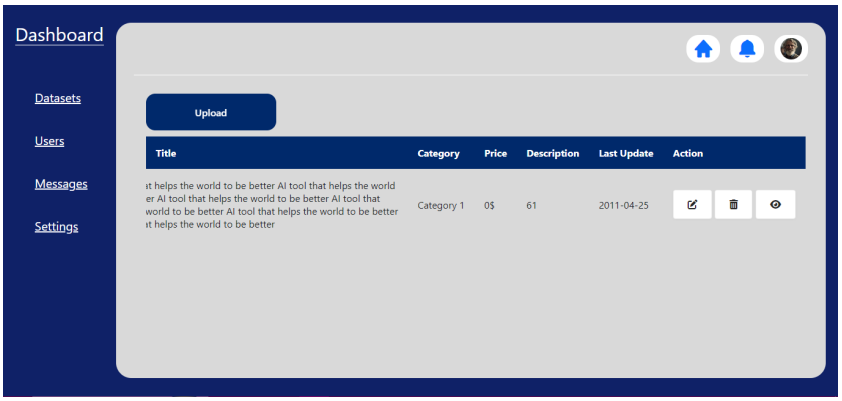


Figure 5.10: Page of Admin dashboard : Datasets Table

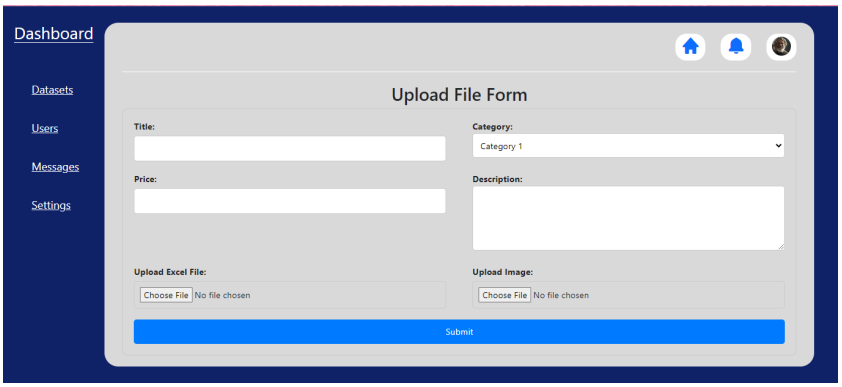


Figure 5.11: Page of Admin dashboard to upload Dataset

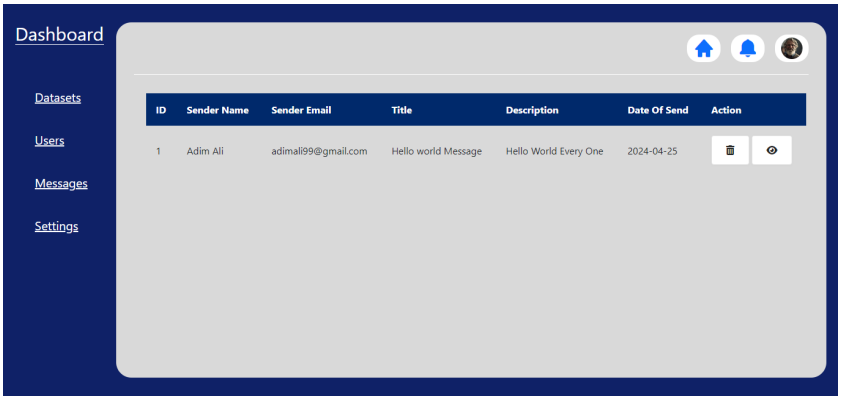


Figure 5.12: Page of Admin dashboard to view messages

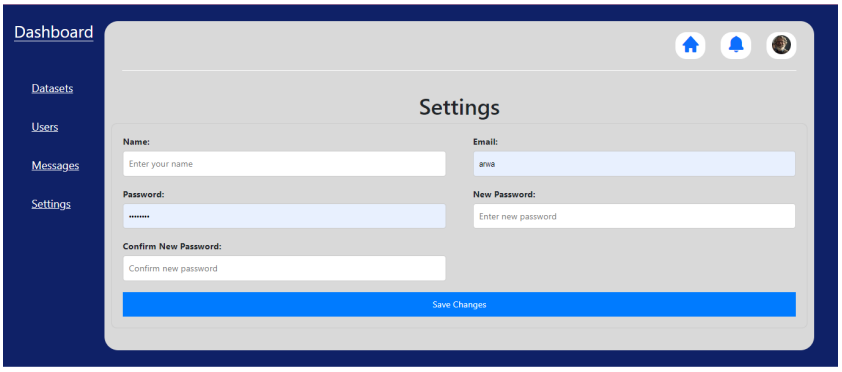


Figure 5.13: Setting Admin dashboard

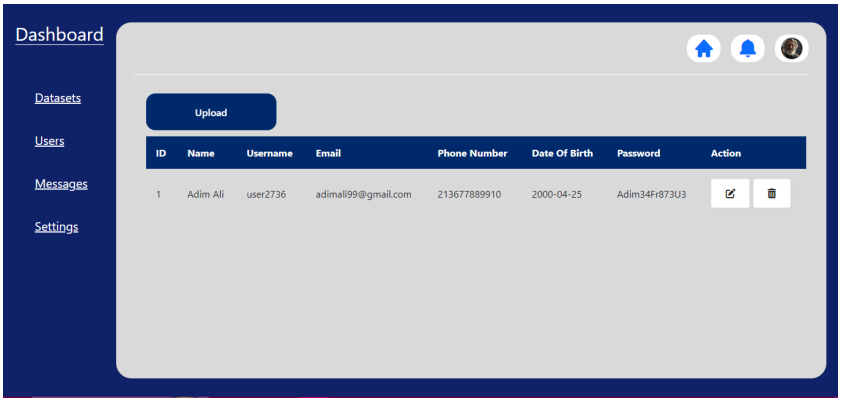


Figure 5.14: Admin dashboard for user management

- **About Us**

This page highlights the company's mission to democratize access to high-quality datasets for AI development, with a special emphasis on Arabic-language datasets. It features an engaging Swiper-powered introduction that outlines the platform's offerings, such as diverse datasets for various AI applications, and the importance of Arabic content. The page also showcases the team members, each dedicated to advancing technology and problem-solving, and details the company's mission, vision, and core values of accessibility, diversity, community, improvement, innovation, and empowerment. This comprehensive page underscores DataBazaar's commitment to fostering collaboration and innovation in the AI community 5.15.

- **Contact Us**

A "Contact Us" page is essential for any website as it serves as a direct communication link between the organization and its users. It enhances customer support by providing a way for users to resolve issues and make inquiries, thus improving satisfaction. The page builds trust and credibility, demonstrating the organization's openness and valuing user feedback.


Figure 5.15: About Us page

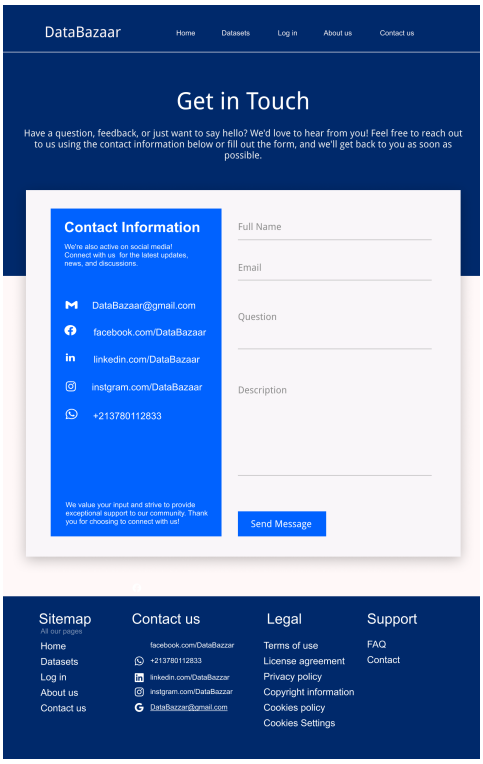


Figure 5.16: Contact Us page

- **Data set Description**

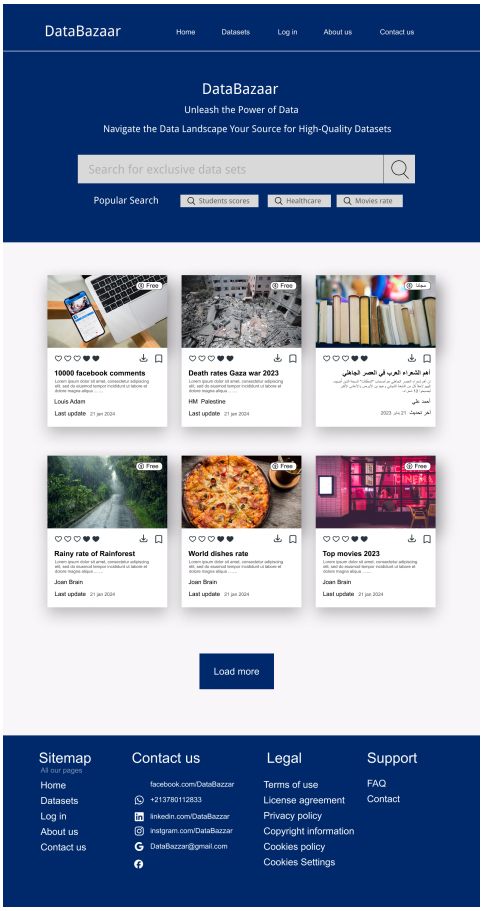


Figure 5.17: Data set page

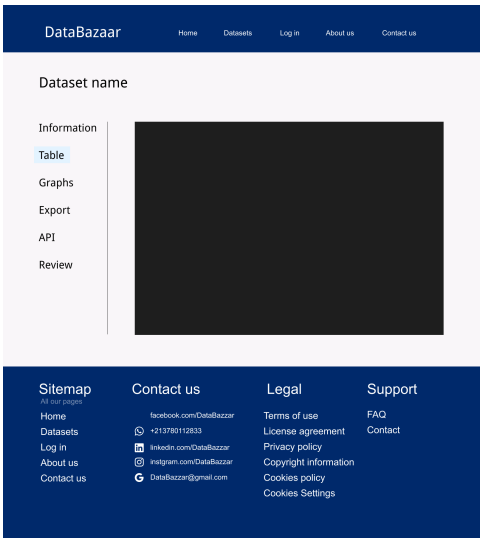


Figure 5.18: Data set Details

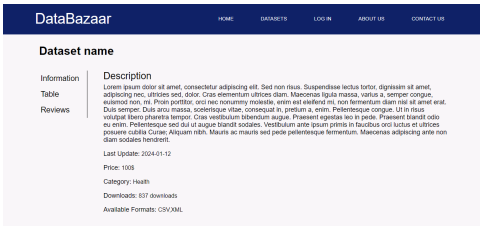


Figure 5.19: Data set Information

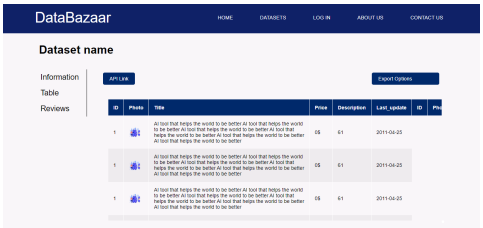


Figure 5.20: Data set Table

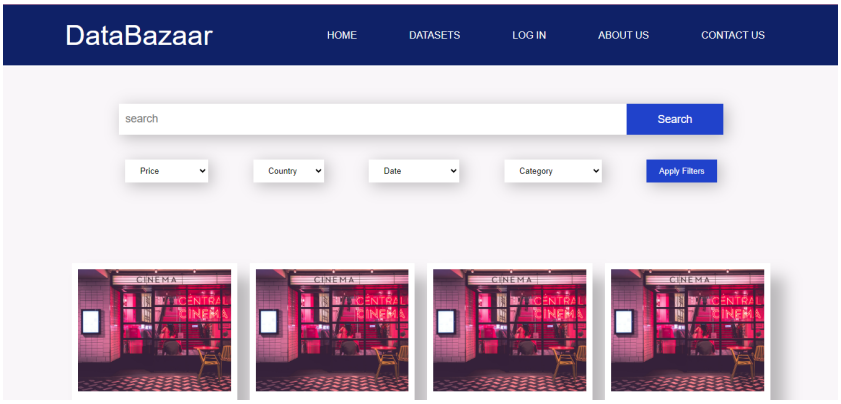


Figure 5.21: Data set Search

### 5.3.2 User Experience and User Interface (UX/UI)

- **Responsivity:** in web design refers to the capacity of a website to adapt its layout and functionality seamlessly across various devices and screen sizes. This adaptation is achieved through techniques like fluid grid layouts, flexible images, media queries, and responsive typography, which allow elements to adjust proportionally to different viewing environments. The importance of responsivity lies in its ability to provide an optimal user experience regardless of the device being used. With the increasing prevalence of mobile browsing, having a responsive website is crucial for attracting and retaining users, as it ensures easy navigation, readability, and interaction on smartphones and tablets. Moreover, responsive design can lead to higher search engine rankings, lower maintenance costs, and future-proofing against emerging technologies and device sizes, making it a fundamental aspect of modern web development.

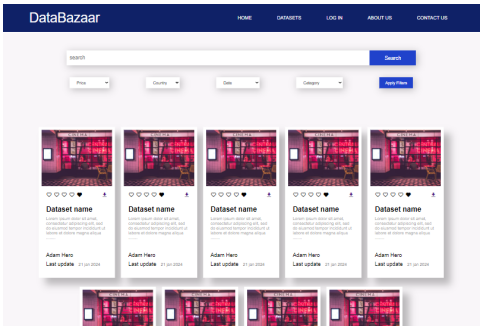


Figure 5.22: Platform view in computer screen

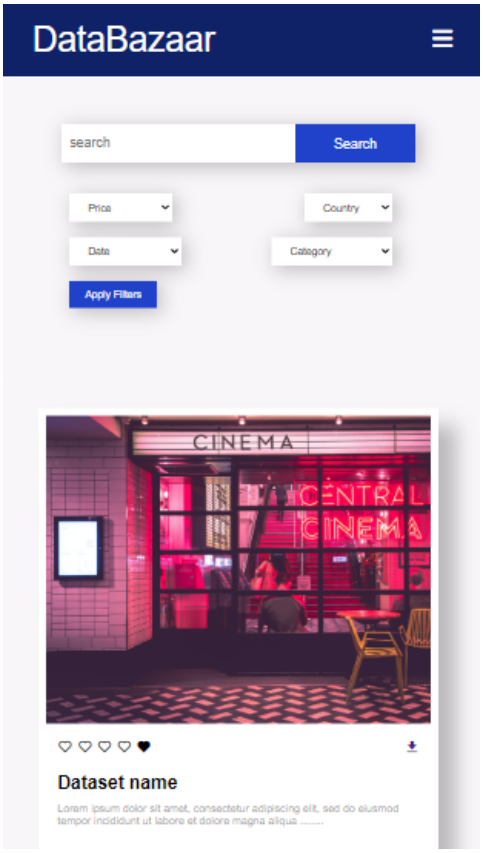


Figure 5.23: Platform view in mobile screen

- **Accessibility:** Multilingual websites offer content in multiple languages, facilitating access for users across diverse linguistic backgrounds. In the case of a dataset platform focused on AI-oriented Arabic content, offering both Arabic and English languages is crucial for enhancing user experience. This feature ensures inclusivity, enabling Arabic-speaking researchers, developers, and enthusiasts to engage with the platform effortlessly in their native language while also catering to an international audience proficient in English. By providing content in multiple languages, the platform promotes accessibility, encourages engagement, and fosters a more user-friendly experience, ultimately facilitating seamless interaction and knowledge exchange within the AI community..


Figure 5.24: Data set description in English language

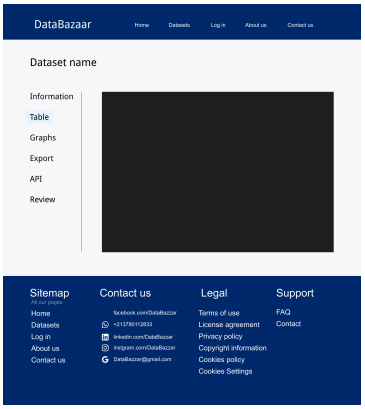


Figure 5.25: Data set description page in English language


Figure 5.26: About US page in Arabic language


Figure 5.27: About US page page in English language

- **Visibility:** Search Engine Optimization (SEO) involves various strategies to improve a website's visibility and ranking in search engine results pages (SERPs).

Keywords and descriptions in meta tags are essential components of SEO.By incorporating relevant keywords into these tags, the platform enhances its visibility in search engine results, attracting more organic traffic. The meta title serves as the clickable headline, while the meta description provides a brief summary of the page's content. Optimizing these tags ensures that search engines accurately understand and represent the platform's content to users, increasing click-through rates and overall user engagement. As such, meta tags play a crucial role in improving the platform's search engine rankings and facilitating user discovery, ultimately contributing to its success and online presence.

```
<meta name="description" content="Our platform serves as a dynamic marketplace,
    offering a vast array of datasets available for both free access and purchase,
    catering to a diverse range of AI applications. Whether you're delving into machine learning,
    natural language processing, computer vision, or any other AI domain.">
<meta name="keywords" content="Arabic Datasets, Datasets, Arabic, datsets for AI, AI, CSV">
```

Figure 5.28: Example of using Meta tags to enhance SEO of the platform

## 5.4 Summary

This chapter outlines the various Implementation Tools essential for development, including Graphic Design Tools, Programming Tools, and other Additional Tools, which are elaborated upon in Sections 5.2.1, 5.2.2, and 5.2.3 respectively. The chapter then progresses to the DataBaazar development in Section 5.3, detailing the creation of Platform Pages and focusing on enhancing the User Experience and User Interface (UX/UI). This comprehensive overview ensures a thorough understanding of the technical and creative processes involved in the successful realization of the platform.

# Conclusion

In this thesis, we have delved into the crucial role that datasets play in the advancement of artificial intelligence, particularly emphasizing the significant gap in resources for Arabic content. Our exploration began with a comprehensive overview of web-based systems and the fundamentals of artificial intelligence, establishing a foundation for understanding the critical importance of high-quality datasets. We identified the disparities in data set availability across different regions and languages, underscoring the challenges faced by the Arabic-speaking world in leveraging AI technologies effectively.

The literature review highlighted the current state of data set platforms, with a specific focus on Arabic-oriented content. We discussed the pressing need for accessible, comprehensive datasets to drive AI applications tailored to Arabic-speaking populations. By examining existing platforms and their regional classifications, we pinpointed the unique obstacles encountered in the identification and utilization of Arabic datasets.

The planning phase delineated the essential objectives, requirements, and business models necessary for creating a sustainable and user-centric platform. In the conception phase, we meticulously outlined the front-end and back-end designs, ensuring a seamless integration of user experience (UX) and user interface (UI) principles with robust platform architecture and database management. The implementation and development of the DataBaazar platform exemplified our practical efforts to enhance the accessibility and quality of Arabic datasets. The implementation and development phase translated these designs into a functional platform, utilizing a variety of graphic design and programming tools to create an efficient and scalable solution. The development of the DataBaazar platform exemplified the practical application of theoretical concepts, culminating in a product that addresses the needs of programmers, AI developers, educational institutions, companies, and government agencies.

In conclusion, this thesis contributes to the body of knowledge by illuminating the existing gaps in AI dataset resources for Arabic content and proposing viable solutions to bridge these gaps. By developing and refining the DataBaazar platform, we aim to empower developers, researchers, and users within the Arabic-speaking community, fostering innovation and inclusivity. The successful implementation of such platforms can drive the growth of AI applications in the Arabic-speaking world, unlocking the potential of AI technologies to benefit a broader, more diverse user base.

Future work should continue to focus on expanding and enriching the datasets available for Arabic AI applications, ensuring that they are comprehensive, representative,

and easily accessible. By doing so, we can further support the integration of AI into various sectors, ultimately contributing to the global advancement of artificial intelligence and its equitable distribution across different linguistic and cultural landscapes.

# Bibliography


[1] *Amazon web services (aws) public datasets*.
https://registry.opendata.aws/.

[2] *Arabic Emotion Recognition*.
https://datasetsearch.research.google.com/search?query=arabic%20emotion%20recognition&docid=TU0eYH7tFCuE8oRcAAAAAA%3D%3D.

[3] *Arabic Handwritten Digits*.
https://www.kaggle.com/mloey1/ahdd1.

[4] *Arabic Image Captioning*.
https://datasetsearch.research.google.com/search?query=arabic%20image%20captioning&docid=2nMlJZtSHGZa8uIsAAAAAA%3D%3D.

[5] *Arabic Named Entity Recognition*.
https://datasetsearch.research.google.com/search?query=arabic%20named%20entity%20recognition&docid=e6DL0%2FARrJ%2Fg%2FOzJAAAAAA%3D%3D.

[6] *Arabic News Text Classification*.
https://www.kaggle.com/nasserali/arabic-news-text-classification.

[7] *Arabic Online Shopping Reviews*.
https://www.kaggle.com/arashnic/arabic-online-shopping-reviews.

[8] *Arabic Poetry Dataset*.
https://www.kaggle.com/alihassan/a-poetry-dataset.

[9] *Arabic Sentiment Twitter Corpus*.
https://huggingface.co/datasets/arabic_sentiment.

[10] *Arabic Speech Recognition*.

https://datasetsearch.research.google.com/search?query=arabic%20speech%20recognition&docid=x_6VXE7ftBjz7VMeAAAAAA%3D%3D.

[11] *Arabic Wikipedia Text*.
https://huggingface.co/datasets/wiki_lingua/ar.

[12] *AraVec: Arabic Word Embeddings*.
https://huggingface.co/sfarah/aravec.

[13] *Coco (common objects in context)*.
http://cocodataset.org/.

[14] *Fast.ai datasets*.
https://course.fast.ai/datasets.

[15] *Github*.
https://github.com/.

[16] *Google ai datasets*.
https://ai.google/tools/datasets/.

[17] *Google dataset search*.
https://datasetsearch.research.google.com/.

[18] *Hugging face datasets*.
https://huggingface.co/datasets.

[19] *Imagenet*.
http://www.image-net.org/.

[20] *Kaggle*.
www.kaggle.com.

[21] *Machine translation of arabic (mtra) project*.
Accessed 2024-05-13.

[22] *MADAR Arabic Dialect Corpus*.
https://huggingface.co/datasets/madar.

[23] *merriam webster  dictionnary*.
https://www.merriam-webster.com/dictionary/dataset.
Accessed: 2024-02-15.

[24] *Microsoft research open data*.
https://msropendata.com/.

[25] *Openai gym*.
https://gym.openai.com/.

[26] *Stanford question answering dataset (squad)*.
https://rajpurkar.github.io/SQuAD-explorer/.

[27] *UCI Machine Learning Repository*.
archive.ics.uci.edu/ml.

[28] *World data info*.
https://www.worlddata.info/languages/arabic.php.
Accessed: 2024-03-04.

[29] M. ABBAS, B. DAHER, AND W. EL-HAJJ, *Arabic search engine: State of the art and challenges*, Journal of King Saud University - Computer and Information Sciences, 29 (2017), pp. 508–520.

[30] M. A. AL-ZOUBE AND M. NIJIM, *E-learning in arabic: Challenges and future directions*, International Journal of Advanced Computer Science and Applications, 8 (2017), pp. 303–309.

[31] M. G. ALOMARI AND Y. M. KADAH, *Ai-driven healthcare applications for arabic-speaking populations*, International Journal of Computer Applications, 181 (2018), pp. 27–34.

[32] D. BROWN AND S.-J. KIM, *Computer vision: Algorithms and applications*, Image Processing Journal, 15 (2020), pp. 211–225.

[33] J. M. CHAQUET, E. J. CARMONA, AND A. FERNÁNDEZ-CABALLERO, *A survey of video datasets for human action and activity recognition*, Computer Vision and Image Understanding, 117 (2013), pp. 633–659.

[34] L. CHEN AND M. JOHNSON, *Robotics: State-of-the-art and future directions*, International Journal of Robotics, 12 (2019), pp. 78–90.

[35] F. Z. CHERFAOUI, A. ZEGNOUN, AND B. ABDULRAZAK, *Content recommendation systems for arabic users: A review*, International Journal of Computer Applications, 169 (2017), pp. 1–5.

[36] K. DAS, J. SCHNEIDER, AND D. B. NEILL, *Anomaly pattern detection in categorical datasets*, in Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, 2008, pp. 169–176.

[37] DRAW.IO, *Draw.io*, 2024.
Accessed: January 14, 2023.

[38] M. ELARABY AND K. DARWISH, *mgpt: An arabic pre-trained generative language model*, in Proceedings of the 12th Language Resources and Evaluation Conference (LREC), European Language Resources Association (ELRA), 2020, pp. 3652–3658.

[39] O. ELNAGAR AND E. EL-SAADAWI, *Maqa: A large scale corpus for arabic healthcare question answering*, in Communications in Computer and Information Science, vol. 625, Springer, 2016, pp. 131–142.

[40] A. ELOUARDIGHI, F. BARIGOU, AND C. CHERKAOUI, *Developing arabic chatbots: Challenges and opportunities*, in Proceedings of the 11th International Conference on Agents and Artificial Intelligence (ICAART), SCITEPRESS - Science and Technology Publications, 2019, pp. 85–93.

[41] P. ESLING AND C. AGON, *Time-series data mining*, ACM Computing Surveys (CSUR), 45 (2012), pp. 1–34.

[42] FIGMA, *Figma*, 2024.
Accessed: April 03, 2024.

[43] M. GARCIA AND R. PATEL, *Natural language processing: Advances and challenges*, Computational Linguistics Review, 28 (2021), pp. 123–138.

[44] N. HABASH, H. BOUAMOR, H. HAJJ, W. MAGDY, W. ZAGHOUANI, F. BOUGARES, N. TOMEH, I. A. FARHA, AND S. TOUILEB, *Proceedings of the sixth arabic natural language processing workshop*, in Proceedings of the Sixth Arabic Natural Language Processing Workshop, 2021.

[45] N. HABASH AND O. RAMBOW, *Arabic natural language processing: Challenges and solutions*, ACM Transactions on Asian and Low-Resource Language Information Processing, 10 (2011), pp. 20:1–20:30.

[46] J. HAN, M. KAMBER, AND J. PEI, *Data mining concepts and techniques third edition*, University of Illinois at Urbana-Champaign Micheline Kamber Jian Pei Simon Fraser University, (2012).

[47] S. S. KUMAR, S. T. AHMED, P. VIGNESHWARAN, H. SANDEEP, AND H. M. SINGH, *Two phase cluster validation approach towards measuring cluster quality in unstructured and structured numerical datasets*, Journal of Ambient Intelligence and Humanized Computing, 12 (2021), pp. 7581–7594.

[48] M. MADDOX, D. GOEHRING, A. J. ELMORE, S. MADDEN, A. PARAMESWARAN, AND A. DESHPANDE, *Decibel: The relational dataset branching system*, in Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases, vol. 9, NIH Public Access, 2016, p. 624.

[49] D. MAKRIS, K. L. KERMANIDIS, AND I. KARYDIS, *The greek audio dataset*, in Artificial Intelligence Applications and Innovations: AIAI 2014 Workshops: CoPA, MHDW, IIVC, and MT4BD, Rhodes, Greece, September 19-21, 2014. Proceedings 10, Springer, 2014, pp. 165–173.

[50] PLANTUML, *Plantuml*, 2024.
Accessed: May 05, 2024.

[51] M. RERABEK AND T. EBRAHIMI, *New light field image dataset*, in 8th International Conference on Quality of Multimedia Experience (QoMEX), 2016.

[52] M. ROBERTS AND P. GUPTA, *Autonomous systems: Challenges and opportunities*, Autonomous Agents Journal, 5 (2017), pp. 45–58.

[53] Y. SAMIH AND W. M. SALLOUM, *Arabic website localization: Best practices and challenges*, International Journal of Computer Applications, 169 (2017), pp. 29–33.

[54] A. SMITH AND D. JOHNSON, *Machine learning: Foundations and applications*, Journal of Artificial Intelligence, 37 (2022), pp. 89–104.

[55] K. SRINIVASAN, K. RAMAN, J. CHEN, M. BENDERSKY, AND M. NAJORK, *Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning*, in Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 2443–2449.

[56] A. E. TAHER AND W. F. ABD-ALMAGEED, *Arabic speech recognition: Challenges and solutions*, IEEE Signal Processing Magazine, 30 (2013), pp. 74–82.

[57] P. WANG, *On defining artificial intelligence*, Journal of Artificial General Intelligence, 10 (2019), pp. 1–37.

[58] X. WANG AND J.-H. PARK, *Expert systems: Knowledge-based problem solvers*, Artificial Intelligence Review, 8 (2018), pp. 189–200.

[59] F. ZARAKET AND H. HAJJ, *Ai solutions for arabic government services*, in Proceedings of the International Conference on Computational Linguistics (COLING), International Committee on Computational Linguistics, 2020, pp. 2113–2122.