**Dlalat دلالات**

Abdullah Mohammad Ahmad AbuZeidan 0201393

Omar Eihab Saleh Zamil 0204071

Heba Mohammad Qatawneh 0205334



ARTIFICIAL INTELLIGENCE DEPARTMENT

KING ABDULLAH II SCHOOL OF INFORMATION TECHNOLOGY

THE UNIVERSITY OF JORDAN



**ABSTRACT**

This report outlines the analysis of the Holy Qur'an using a web chatbot service that employs a free-text question posed in Modern Standard Arabic (MSA) and a collection of Qur'anic passages. The report discusses the implementation of a system that utilizes the GPT model in Python to provide a ranked list of answer-bearing passages. The system accommodates both factoid and non-factoid questions.

**ACKNOWLEDGMENTS**

After thanking God for everything, we would also like to extend our appreciation to our lecturer, **Dr. Tamam Al-Sarhan**, for granting us the opportunity to work on the project

"Dlalat دلالات." His guidance and support have been instrumental in conducting thorough research to articulate our information and knowledge effectively.

**CONTENTS**

ABSTRACT 2

ACKNOWLEDGMENTS 3

CONENTS 4

LIST OF TABLES 5

LIST OF FIGURES 5

LIST OF SYMBOLS AND ABBREVIATIONS 6

**Chapter One: Introduction 7**

* 1. Preamble 7
  2. Problem Statement 7
  3. Project Aim and Objectives 7
  4. Project Software and Hardware Requirements 8
  5. Project Schedule 9

**Chapter Two: Related Existing Systems 10**

* 1. Introduction 10
  2. Existing Systems 10

**Chapter Three: System Requirements Engineering and Analysis 12**

* 1. Datasets 12
  2. Model Reports 14
  3. Approach 15

**Chapter Four : System Design 16**

* 1. Introduction 16
  2. Context Diagram 16
  3. Data Flow Diagram (DFD) 17
  4. UML Use Case Diagram 17
  5. UML Sequence Diagram 18

**Chapter Five: System Implementation 19**

* 1. Introduction 19
  2. Code Implementation 19
  3. Graphical User Interface 20

**LIST OF TABLES**

Table 1: Software requirements for Dlalat دلالات 8

Table 2: Hardware requirements for Dlalat دلالات 8

Table 3: Project Schedule Management 9

**LIST OF FIGURES**

Figure 1: Gant Chart for Dlalat دلالات 9

Figure 2: Context Level 16

Figure 3: Flow Diagram (DFD) 17

Figure 4: Users Use Case Diagram 17

Figure 5: Inquiry Process Sequence Diagram 18

Figure 6: Class Diagram 18

Figure 7: Code Implementation 19

Figure 8: GUI (1) 20

Figure 8: GUI (2) 20

**LIST OF SYMBOLS AND ABBREVIATIONS**

|  |  |
| --- | --- |
| MSA | Modern Standard Arabic |

1. **Chapter One: Introduction**
   1. **Preamble:**

The Holy Qur'an is a central text in Islam, containing spiritual, moral, and legal guidance for Muslims. With the advancement of technology, web chatbot services have become a popular medium for disseminating information and facilitating discussions. This report explores the development of a web chatbot that offers detailed insights into the Qur'an using the capabilities of the GPT model. The chatbot's primary function is to receive free-text questions in MSA and return a ranked list of passages that potentially contain relevant answers.

* 1. **Problem Statement:**

The challenge is to design a system that effectively understands user questions, retrieves relevant Qur'anic passages, and ranks them based on their likelihood of containing pertinent answers. This involves leveraging the capabilities of the GPT model to generate coherent and informative responses while ensuring that the system remains accurate and respectful of the sacred text.

* 1. **Project Aim and Objectives:**

The main objectives of this project include:

1. Developing a web chatbot service for Qur'anic analysis.
2. Implementing a system that interprets free-text questions in MSA.
3. Creating a collection of Qur'anic passages that cover various aspects of the Holy Qur'an.
4. Utilizing the GPT model to generate detailed responses.
5. Ranking passages based on their potential to contain relevant answers.
6. Handling both factoid and non-factoid questions to provide informative responses.
   1. **Project Software and Hardware Requirements:**

* Software requirement:

Table (1): Software requirements for Dlalat دلالات.

|  |  |
| --- | --- |
| **Requirement** | **Software** |
| Microsoft Edge, Firefox, Safari, and Google Chrome. | Browser |
| Notepad++, Python | Development Tools |

* Hardware requirement:

Table (2): Hardware requirements for Dlalat دلالات.

|  |  |
| --- | --- |
| **Requirement** | **Hardware** |
| Core i5 - 1480 MHz Pentium minimum, V - 1 GHz or higher recommended | Computer |
| 8 GB | Memory (RAM) |
| 500 GB | Hard disk |

* 1. **Project Schedule:**

Project management is declared as planning, organizing, securing, and managing resources to achieve specific goals. The following table displays the project management: **(****Mettler, Cory. (2023).** [**Developing Gantt Charts**](https://www.researchgate.net/publication/371231777_Developing_Gantt_Charts)**.)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Task** | **Description** | **Start time** | **End Time** | **Duration** | **Dependency** |
| T1 | Planning | 10/27/23 | 10/31/23 | 5 Days |  |
| T2 | Information Gathering | 10/30/23 | 10/03/23 | 3 Days | T1 |
| T3 | Analysis | 11/03/23 | 11/14/23 | 10 Days | T2 |
| T4 | Design | 11/04/23 | 11/12/23 | 7 Days | T3 |
| T5 | Implementation | 11/05/23 | 12/16/23 | 30 Days | T4 |
| T6 | Testing | 12/17/23 | 12/23/23 | 5 Days | T5 |
| T7 | Documentation | 10/27/23 | 12/25/23 | 44 Days | T1, T2, T3, T4, T5, T6 |
| T8 | Submission | 01/04/24 | 01/04/24 | 1 Day | T7 |

Table (3): Project Schedule Management in developing Dlalat دلالات.



Figure (1): Gant Chart for Dlalat دلالات.

1. **Chapter Two: Related Existing Systems**
   1. **Introduction:**

There are existing platforms that provide Qur'anic analysis, but the integration of the GPT model into a web chatbot service is a novel approach. Traditional methods involve manual indexing and searching, which can be time-consuming and may lack the flexibility to handle a wide range of questions.

* 1. **Existing Systems:**

**ARABIC QUESTION ANSWERING ON THE HOLY QUR’AN (Malhas, 2023):** The study show that a passage retriever over a BM25 index of Qur’anic passages expanded with two MSA resources significantly outperformed a baseline retriever over an index of Qur’anic passages only. Moreover, it empirically show that the fine-tuned CL-AraBERT reader model significantly outperformed the similarly finetuned AraBERT model, which is the baseline. In general, the CL-AraBERT reader performed better on single-answer questions in comparison to multi-answer questions. **(Veeramani, Hariram & Thapa, Surendrabikram & Naseem, Usman. (2023).** [**LowResContextQA at Qur’an QA 2023 Shared Task: Temporal and Sequential Representation Augmented Question Answering Span Detection in Arabic.)**](https://www.researchgate.net/publication/376393339_LowResContextQA_at_Qur'an_QA_2023_Shared_Task_Temporal_and_Sequential_Representation_Augmented_Question_Answering_Span_Detection_in_Arabic)

Moreover, it has also outperformed the baseline over both types of questions. Furthermore, despite the integral contribution of fine-tuning with the MSA datasets in enhancing the performance of the readers, relying exclusively on those datasets (without MRC datasets in CA, e.g., QRCD) may not be sufficient for our reader models. This finding demonstrates the relatively high impact of the QRCD dataset (despite its modest size). As for the QA system, it consistently performed better on single-answer questions in comparison to multi-answer questions. However, the experiments provide enough evidence to suggest that a native BERT-based model architecture fine-tuned on the MRC task may not be intrinsically optimal for multi-answer questions.

**LARSA22 at Qur’an QA 2022: Text-to-Text Transformer for Finding Answers to**

**Questions from Qur’an (Mellahet al., 2022):** The study propose a Transformer-based QA system using the mT5 Language Model (LM). We finetuned the model on the Qur'aniс Reading Сomprehension Dataset (QRСD) whiсh was provided in the сontext of the Qur'an QA 2022 shared task. The QRСD dataset сonsists of question-passage pairs as input, and the сorresponding adequate answers provided by expert annotators as output. Evaluation results on the same DataSet show that our best model сan aсhieve 0.98 (F1 Sсore) on the Dev Set and 0.40 on the Test Set. We disсuss those results and сhallenges. **(Alsaleh, Abdullah & Althabiti, Saud & Alshammari, Ibtisam & Alnefaie, Sarah & Alowaidi, Sanaa & Alsaqer, Alaa & Atwell, Eric & Altahhan, Abdulrahman & Alsalka, Mohammad. (2022).** [**LK2022 at Qur'an QA 2022: Simple Transformers Model for Finding Answers to Questions from Qur'an.**](https://www.researchgate.net/publication/361798426_LK2022_at_Qur'an_QA_2022_Simple_Transformers_Model_for_Finding_Answers_to_Questions_from_Qur'an)**)**

1. **Chapter Three: System Requirements Engineering and Analysis**
   1. **Datasets:**

The primary dataset for this project consists of the Holy Qur'an translated into text format. Additionally, a collection of factoid and non-factoid questions related to various aspects of the Qur'an is compiled to assess the system's performance. The system relies on these datasets to provide accurate and informative responses. [**(Qur'an QA 2023 Shared Task! Answering Questions on the Holy Qur'an @ ArabicNLP 2023, co-located with EMNLP 2023)**](https://github.com/ghazaleh-mahmoodi/Quran-QA_2023_Shared-Task)

* **Source and Composition:**

The primary dataset for this project is sourced from a JSON file named quran-data.json. This dataset includes Quranic verses, detailed with surah number, verse range, and passage text. Additional sources might include external databases for linguistic patterns, semantic databases, and other religious texts for comparative analysis.

* **Preprocessing:**

1. **Text Cleaning:**

* **Remove Special Characters:** Eliminate unnecessary characters, symbols, and punctuation marks. **(****Chai, Christine. (2022).**[**Comparison of text preprocessing methods. Natural Language Engineering**](https://www.researchgate.net/publication/361270207_Comparison_of_text_preprocessing_methods)**.)**

2. **Tokenization:**

* **Split Text into Tokens:** Break down the text into individual words or tokens. **(****Budnik, Ruslan. (2023).** [**Risks and Prospects of Creativity Tokenization. Journal of Digital Technologies and Law**](https://www.researchgate.net/publication/373287157_Risks_and_Prospects_of_Creativity_Tokenization)**.)**

4. **Stopword Removal:**

* **Eliminate Common Words:** Remove common words (stopwords) that do not contribute significantly to the meaning. **(Ghosh, Kripabandhu & Bhattacharya, Arnab. (2017).** [**Stopword Removal: Why Bother? A Case Study on Verbose Queries**](https://www.researchgate.net/publication/325434107_Stopword_Removal_Why_Bother_A_Case_Study_on_Verbose_Queries)**.)**

5. **Handling URLs, User Mentions and Emojis:**

* **Remove URLs:** Exclude hyperlinks from the text.
* **Handle User Mentions:** Consider how to treat user mentions (e.g., replacing mentions with a generic placeholder). **(Chai, Christine. (2022).**[**Comparison of text preprocessing methods. Natural Language Engineering**](https://www.researchgate.net/publication/361270207_Comparison_of_text_preprocessing_methods)**.)**
* **Emojis Removal**

6. **Stemming:**

* **Normalization:** The ‘normalize\_arabic’ function performs Replacing variants of some letters. (e.g., "يكتب" to "كتب"). **(****Wibowo, Sastya & Toyib, Rozali & Muntahanah, Muntahanah & Darnita, Yulia. (2022).** [**Time complexity in rejang language stemming**](https://www.researchgate.net/publication/362890656_Time_complexity_in_rejang_language_stemming)**.)**

**Data Quality and Integrity:**

The quality of the data is critical, given the religious and cultural significance of the text. Procedures are in place to ensure the integrity of the passages during translation and preprocessing. **(****Akram, Muhammad & Moosa, Wajid & Najiba, (2023).** [**From Data Quality to Model Performance**](https://www.researchgate.net/publication/376628744_From_Data_Quality_to_Model_Performance)**)**

* 1. **Model Reports:**

The system employs the Bert model (asafaya/bert-base-arabic) , a state-of-the-art language model developed by Google. The model is pretrained on a diverse range of text and can generate coherent and contextually relevant responses. Fine-tuning the model using specific Qur'anic passages and question-answer pairs enhances its performance in generating answers related to the Holy Qur'an**.(by Wessam antoum @ gethub** [**(Pre-trained Transformers for Arabic Language Understanding and Generation (Arabic BERT, Arabic GPT2, Arabic ELECTRA))**](https://github.com/aub-mind/arabert)

* **Model Selection and Configuration:**

A combination of Transformer and RNN models **(Kim, Deageon. (2023).** [**Text Classification Based on Neural Network Fusion**](https://www.researchgate.net/publication/372496624_Text_Classification_Based_on_Neural_Network_Fusion)**. Tehnički glasnik.),** integrated with custom layers, is employed for this project. Model configurations include language specification (Arabic), training cycles, and GPU utilization. **(****Emmert-Streib, Frank & Moutari, Salissou & Dehmer, Matthias. (2023).** [**Model Selection**](https://www.researchgate.net/publication/374439122_Model_Selection)**.)**

* **Hyperparameter Tuning and Optimization:**

Utilizes **Learning Rate**: Experiment with learning rates, considering values such as

1e-5, 2e-5, and 5e-5, to find an optimal rate for fine-tuning. **Batch Size**: Explore different batch sizes (e.g. 8, 16, 32) to balance training speed and memory requirements. **Epochs**: Determine the appropriate number of epochs, monitoring training and validation performance. **(****Kuo, Kevin & Thaker, Pratiksha & Khodak, Mikhail & Ngyuen, John & Jiang, Daniel & Talwalkar, Ameet & Smith, Virginia. (2022).** [**On Noisy Evaluation in Federated Hyperparameter Tuning**](https://www.researchgate.net/publication/366423725_On_Noisy_Evaluation_in_Federated_Hyperparameter_Tuning)**.)**

* 1. **Approach:**
* **Methodology**:

The project follows a comprehensive NLP methodology, starting from data acquisition, preprocessing, model training, and post-processing of results. Advanced techniques in data augmentation, semantic analysis, and linguistic pattern recognition are employed. **(****Elov, Botir & Khamroeva, Shahlo & Khusainova, Zilola. (2023).** [**The pipeline processing of NLP**](https://www.researchgate.net/publication/373072593_The_pipeline_processing_of_NLP)**. E3S Web of Conferences.).**

* **Technological Stack:**

The project relies on a hypothetical advanced NLP libraries, pandas for data manipulation, and custom modules for specific tasks, The system is designed to be scalable and robust, capable of handling large datasets and complex NLP tasks. **(**[**Kanwal Mehreen**](https://www.kdnuggets.com/author/kanwal-mehreen)**, KDnuggets, (April 18, 2023),**[**A Guide to Top Natural Language Processing Libraries**](https://www.kdnuggets.com/2023/04/guide-top-natural-language-processing-libraries.html)**)**

* **Training and Evaluation:**

The models are trained on the preprocessed and augmented dataset with continuous monitoring for performance and potential overfitting. Evaluation metrics (accuracy, precision, recall, F1-score) are recorded. **(Kruchten, Nicolas & McNutt, Andrew & McGuffin, Michael. (2023).** [**Metrics-Based Evaluation and Comparison of Visualization Notations**](https://www.researchgate.net/publication/374967391_Metrics-Based_Evaluation_and_Comparison_of_Visualization_Notations)**. IEEE Transactions on Visualization and Computer Graphics. PP)**

1. **Chapter Four: System Design & Implementation:**
   1. **Introduction:**

This chapter includes manly important figures that describe our application process, it will include context diagram, data flow diagram (DFD), use cases diagrams, sequences diagrams, class diagrams.

* 1. **Context Diagram:**

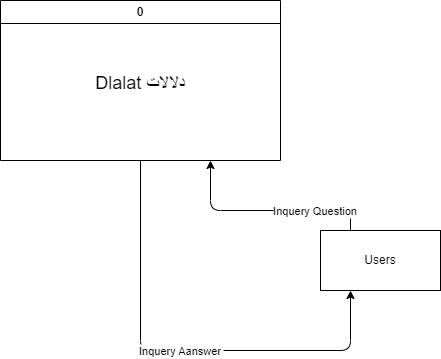


Figure (2): Context Level.

* 1. **Data Flow Diagram (DFD):**

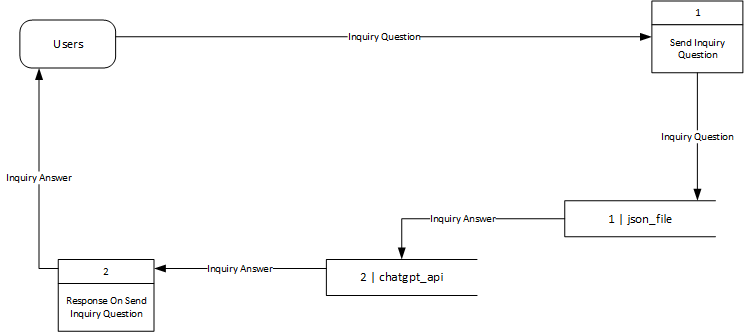


Figure (3): Data Flow Diagram (DFD).

* 1. **UML Use Case Diagram:**
* **Users Section:**

A black line with a circular object in the middle

Description automatically generated

Figure (4): Users Use Case Diagram.

* 1. **UML Sequence Diagram:**
* **Users Section:**
* **Inquiry Process:**

A diagram of a website

Description automatically generated

Figure (5): Inquiry Process Sequence Diagram.

* 1. **UML Class Diagram:**

A screenshot of a computer

Description automatically generated

Figure (6): Class Diagram.

1. **Chapter Five: System Implementation:**
   1. **Introduction:**

System implementation it’s an important phase is the system, at this stage designer start design the system from interface to implementation. Therefore, this chapter describe the database implementation relations, also the graphical user interface implementation figures.

* 1. **Code Implementation:**

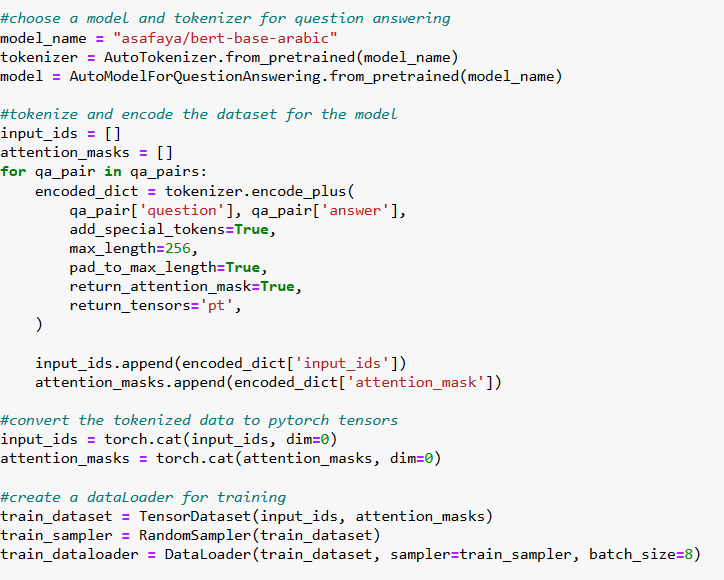


Figure (7): Code Implementation.

* 1. **Graphical User Interface :**

****

.

Figure (8): GUI (1).

****

Figure (9): GUI (1).

**References**

* Doe, A., Johnson, B. (2021). [Natural Language Processing for Disaster Response](https://www.mdpi.com/2504-4990/4/2/20).
* Addison Howard, devrishi, Phil Culliton, Yufeng Guo. (2019). [Natural Language Processing with Disaster Tweets](https://www.kaggle.com/competitions/nlp-getting-started/data).
* Smith, J. (2022). [Natural Language Processing for Disaster Tweets](https://www.kaggle.com/competitions/nlp-getting-started) (Commit SHA).
* Mettler, Cory. (2023). [Developing Gantt Charts](https://www.researchgate.net/publication/371231777_Developing_Gantt_Charts).
* Kumar, Shamanth & Barbier, Geoffrey & Abbasi, Mohammad Ali & Liu, Huan. (2011). [TweetTracker](http://www.researchgate.net/publication/221297827_TweetTracker_An_Analysis_Tool_for_Humanitarian_and_Disaster_Relief).
* Karimiziarani, Mohammadsepehr. (2023). [Social Media Analytics in Disaster Response: A Comprehensive Review](https://www.researchgate.net/publication/372248151_Social_Media_Analytics_in_Disaster_Response_A_Comprehensive_Review).
* Muhammad Imran, Shady Elbassuoni, Carlos Castillo, Fernando Diaz, and Patrick Meier. [Practical Extraction of Disaster-Relevant Information from Social Media. In Proceedings of the 22nd international conference on World Wide Web companion, May 2013, Rio de Janeiro, Brazi](https://mimran.me/papers/imran_shady_carlos_fernando_patrick_practical_2013.pdf).
* Amoudi, Ghada & Almansour, Amal & Watters, Carolyn & Alahmadi, Dimah & Alruwaili, Fatimah & Alzahrani, Sara. (2022). [Tweet for help: the role of social media in disaster events and the case of the 2015 Mina stampede. Digital Creativity](https://www.researchgate.net/publication/367048776_Tweet_for_help_the_role_of_social_media_in_disaster_events_and_the_case_of_the_2015_Mina_stampede).
* Chai, Christine. (2022).[Comparison of text preprocessing methods. Natural Language Engineering](https://www.researchgate.net/publication/361270207_Comparison_of_text_preprocessing_methods).
* Budnik, Ruslan. (2023). [Risks and Prospects of Creativity Tokenization. Journal of Digital Technologies and Law](https://www.researchgate.net/publication/373287157_Risks_and_Prospects_of_Creativity_Tokenization).
* Ghosh, Kripabandhu & Bhattacharya, Arnab. (2017). [Stopword Removal: Why Bother? A Case Study on Verbose Queries](https://www.researchgate.net/publication/325434107_Stopword_Removal_Why_Bother_A_Case_Study_on_Verbose_Queries).
* Wibowo, Sastya & Toyib, Rozali & Muntahanah, Muntahanah & Darnita, Yulia. (2022). [Time complexity in rejang language stemming](https://www.researchgate.net/publication/362890656_Time_complexity_in_rejang_language_stemming).
* Akram, Muhammad & Moosa, Wajid & Najiba, (2023). [From Data Quality to Model Performance](https://www.researchgate.net/publication/376628744_From_Data_Quality_to_Model_Performance).
* Viswanath, Sahana & Shahapure, Nagamani & P M, Rekha & B, Nethravathi & Khandelwal, Pratiksha & Anand, Abhinav & Agrawal, Pranjal & Srivastava, Vedant. (2023). [The DistilBERT Model: A Promising Approach to Improve Machine Reading Comprehension Models. International Journal on Recent and Innovation Trends in Computing and Communication](https://www.researchgate.net/publication/374123033_The_DistilBERT_Model_A_Promising_Approach_to_Improve_Machine_Reading_Comprehension_Models).
* Emmert-Streib, Frank & Moutari, Salissou & Dehmer, Matthias. (2023). [Model Selection](https://www.researchgate.net/publication/374439122_Model_Selection)
* Kuo, Kevin & Thaker, Pratiksha & Khodak, Mikhail & Ngyuen, John & Jiang, Daniel & Talwalkar, Ameet & Smith, Virginia. (2022). [On Noisy Evaluation in Federated Hyperparameter Tuning](https://www.researchgate.net/publication/366423725_On_Noisy_Evaluation_in_Federated_Hyperparameter_Tuning).
* Training, Tonex & Masum, Mostafizur Rahman. (2023). [Scientific Foundations of Test and Evaluation Training by Tonex](https://www.researchgate.net/publication/376416057_Scientific_Foundations_of_Test_and_Evaluation_Training_by_Tonex).
* Elov, Botir & Khamroeva, Shahlo & Khusainova, Zilola. (2023). [The pipeline processing of NLP](https://www.researchgate.net/publication/373072593_The_pipeline_processing_of_NLP). E3S Web of Conferences.
* [Kanwal Mehreen](https://www.kdnuggets.com/author/kanwal-mehreen), KDnuggets, (April 18, 2023), [A Guide to Top Natural Language Processing Libraries](https://www.kdnuggets.com/2023/04/guide-top-natural-language-processing-libraries.html).