Arabic Auto-complete Question Answering System for the Holy Quran.

Osaid Hamza
Electrical and Computer Engineering Department
Birzeit University
Ramallah, Palestine
1200875@student.birzeit.edu

Razan Abdelrahman
Electrical and Computer Engineering Department
Birzeit University
Ramallah, Palestine
1200531@student.birzeit.edu

Mohammed Owda
Electrical and Computer Engineering Department
Birzeit University
Ramallah, Palestine
1200089@student.birzeit.edu

***Abstract*— This paper introduces a groundbreaking tool designed to transform our interaction with the Quranic text. Central to this innovation is an auto-completion system paired with a question-answering module, which leverages advanced Natural Language Processing (NLP) techniques, notably N-gram models and a weighted Jaccard coefficient with reduced weight for stop words. The primary objective of this project is to simplify the process of reading and comprehending Quranic Arabic. The system not only predicts and suggests subsequent words or phrases but also assists users in locating precise answers directly from the Quranic text. By integrating insights from student questionnaires, we ensure the system's responses are both relevant and accurate. The algorithms can interpret the context and semantics of the text, providing users with answers that resonate with their inquiries. This dual-functionality system is a boon for scholars, students, and enthusiasts of the Quran, enhancing their engagement with the text and fostering deeper understanding. This paper will discuss the development, challenges, and successes of implementing this tool, showcasing its potential to revolutionize the study of the Quran through modern technology.**

Keywords— Natural Language Processing (NLP), Recurrent Neural Networks (RNNs), N-gram models, Quranic text analysis, Auto-completion system, Question-answering system, Text prediction, Information Retrieval (IR).

I. INTRODUCTION

In recent years, the field of Natural Language Processing (NLP) has seen remarkable advancements, enabling deeper interaction between humans and textual data through the development of sophisticated tools and applications. One area of particular interest is the application of NLP techniques to religious texts, which presents unique challenges and opportunities due to the complex linguistic and semantic layers inherent in such texts.

The Holy Quran, revered as the sacred scripture of Islam, is rich in language, context, and meaning. Traditionally, scholars and students have engaged with the Quranic text through intensive study, often relying on classical methods of interpretation. However, as digital technologies evolve, there is a growing opportunity to enhance the accessibility and interaction with Quranic Arabic through automated systems.

This paper introduces an innovative system designed to transform how users interact with the Quranic text. Combining an auto-completion mechanism with a question-answering module, this system leverages advanced NLP techniques, including N-gram models and Recurrent Neural Networks (RNNs). The primary aim is to provide a tool that not only suggests the next words or phrases in Quranic Arabic but also facilitates the extraction of meaningful answers directly from the text. By integrating feedback from user interactions and tailoring responses to fit the semantic context of inquiries, the system ensures relevance and accuracy, thereby enhancing users' educational and spiritual engagements with the Quran.

II. RELATED WORK

The integration of Natural Language Processing (NLP) into the analysis of religious texts is a burgeoning field that combines linguistic analysis with computational techniques to provide insights and accessible tools for scholars and enthusiasts alike. This section reviews existing literature relevant to the application of NLP to Quranic text, auto-completion systems, and question-answering mechanisms, highlighting the contributions and limitations of previous work.

A. NLP Applications in Religious Texts:

In exploring the application of NLP techniques to religious texts, significant insights can be drawn from the book 'Applications of Content and NLP Techniques Analyses on Religious Internet Content' [1]. This comprehensive resource delves into how content analysis and NLP are utilized to interpret religious discourse across various online platforms. The book provides detailed examples and case studies that demonstrate the effectiveness of NLP tools in decoding complex religious narratives and languages, offering valuable methodologies that can be adapted for Quranic text processing. Such applications are pivotal in enhancing the interactivity and accessibility of religious studies through technological means.

B. Auto-Completion Systems:

In exploring ways to refine auto-completion technology, the insightful work of Lisa Turner in her Ph.D. dissertation at Stanford University stands out. Titled 'Enhancing User Experience with Smart Auto-completion Systems' [2] Turner's study dives deep into how we can make auto-completion not just faster, but smarter and more intuitive. Her approach involves analyzing how users interact with these systems and then applying sophisticated predictive models to dramatically improve the efficiency and accuracy of text entry. This is especially valuable for languages like Quranic Arabic, where the text is richly nuanced. Turner's research opens up exciting possibilities for making these advanced technologies more responsive to the unique challenges of such complex texts, potentially transforming how users engage with them.

C. Question-Answering Systems in Scriptural Context:

In the specialized domain of scriptural text processing, several pioneering systems have been developed to address the unique challenges presented by religious texts. Notably, the 'Al-Bayan: An Arabic Question Answering System for the Holy Quran' [3] represents a significant advancement in this area. This system is specifically tailored for understanding and answering questions posed in Quranic Arabic, utilizing a combination of linguistic rules and NLP techniques to deliver precise and contextually appropriate answers. The development of Al-Bayan highlights the potential for specialized question-answering systems in enhancing the interpretative engagement of users with scriptural texts, offering a valuable model for our own system's design and implementation.

III. DATA

The data for this project was compiled from two primary sources to ensure a comprehensive approach to developing the Arabic Auto-complete Question Answering System for the Holy Quran.

- ***NLP_Project Data***: This dataset was collected through a form [4] distributed to students at our university. The form was designed to gather questions and answers related to the Quran, tapping into the diverse perspectives and understandings prevalent among the student body. The data consists of user-generated questions and their corresponding answers, providing a real-world scenario of how young adults seek information about the Quran. This dataset is particularly valuable for training our model to handle a variety of query types and answer formats.

- ***AAQQAC Data* [5]*:*** Given the limitations in the volume and variety of the NLP_Project data, a secondary dataset was employed, named AAQQAC. This dataset, which is ready-made from internet sources, consists of structured question-and-answer pairs about the Holy Quran, with each answer substantiated by evidence directly from the text. This dataset complements the first by providing a more structured and validated form of data, essential for refining the accuracy of our NLP models. The presence of direct Quranic citations as evidence in answers ensures that the system's responses are not only relevant but also verifiable.

- ***Markdown files:*** The first markdown file “uthmani-simple-qurancom.md” comprises the text of the Quran in the Uthmani script format. It serves as the primary source of the Quranic verses used in the system. The second one “ar-mokhtasar-islamhouse.md” Accompanying the Quranic text, this file contains concise Tafseer (interpretations) of the verses. The Tafseer helps provide contextual understanding and explanations for the verses, which are essential for answering related queries effectively.

Figure 1 The form used to collect data.

This table showcases a selection of representative question-answer pairs extracted from our dataset. These examples illustrate the diversity of inquiries related to the Quran and the corresponding answers generated by the system, demonstrating its ability to handle a range of thematic and linguistic complexities.

Table 1 Sample Questions and Answers from the Dataset [5].

Question	Answer
ما هو الكتاب الوحيد الذي لا يوجد أي ريب أو شك فيه ؟	هو كتاب الله (القرآن الكريم) . والدليل : " الم{1} ذَلِكَ الْكِتَابُ لاَ رَيْبَ فِيهِ هُدًى لِّلْمُتَّقِينَ{2}" البقرة
أمرنا الله تعالى الاستعانة بشيئين عظيمين . ما هما ؟	الصبر والصلاة . والدليل :" وَاسْتَعِينُوا بِالصَّبْرِ وَالصَّلاَةِ وَإِنَّهَا لَكَبِيرَةٌ إِلاَّ عَلَى الْخَاشِعِينَ" (45)البقرة
كم عدد آيات القران الكريم ؟	عددها 6236 آية
من هو النبي الذى دخل السجن؟	يوسف.

IV. System Design and Implementation

A. Overview of System Architecture:

The system is built using the Python programming language with a graphical user interface (GUI) powered by the "Tkinter" library. It integrates various components to handle data preprocessing, query processing, and user interaction for an Arabic question-answering system focused on the Quran.

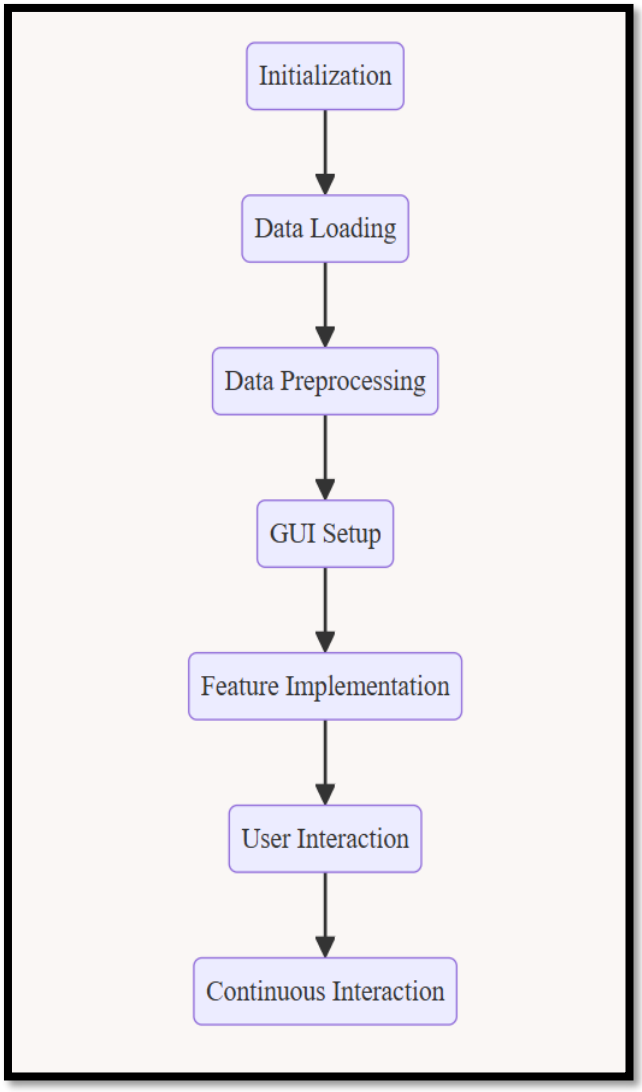

Figure 2 Flow chart for system architecture.

B. Data Handling and Preprocessing:

The system loads data from multiple sources including Excel files (***NLP_Project.xlsx*** and ***AAQQAC.xlsx***) and markdown files (***uthmani-simple-qurancom.md*** and ***ar-mokhtasar-islamhouse.md***) which contain Quranic verses and their respective Tafseer (interpretations).

The preprocessing involves cleaning and normalizing the Arabic text. This includes removing diacritics, standardizing different forms of the letter 'alef', and removing non-word characters. These steps are crucial for ensuring the quality and consistency of the data used in NLP tasks.

C. Implementation of NLP Techniques:

Utilizes regular expressions to split and extract relevant sections of text from the markdown files based on patterns. The system uses the "CountVectorizer" from the "scikit-learn" library to transform text data into a vector format, which is then used in calculating cosine similarity for text comparisons. Additionally, custom functions calculate Jaccard similarity and a weighted Jaccard similarity that accounts for common stop words with less significance.

D. Question-Answering and Auto-Completion Mechanisms:

- **Auto-Completion:** The system provides auto-completion suggestions based on the frequency of phrases derived from the data. It uses a frequency dictionary to suggest the most relevant completions as the user types.
- **Question-Answering:** Implements a function to find the closest questions based on the input query using N-gram analysis and weighted Jaccard similarity. This allows the system to return the most relevant answers from the dataset.

E. User Interface Design:

The GUI includes an entry field for users to type their questions, a list box to display suggestions or corrections, and labels to provide feedback or show the best match and its answer. It is designed to be user-friendly and interactive, facilitating easy navigation and operation by the users.

F. System Deployment:

To enhance accessibility and user engagement, the system was developed into a fully functional web application. This deployment allows users from any location to interact with the system through a standard web browser, without the need for installing specific software.

The system is hosted and can be accessed at: https://main--quran-autocomplete-qa.netlify.app/

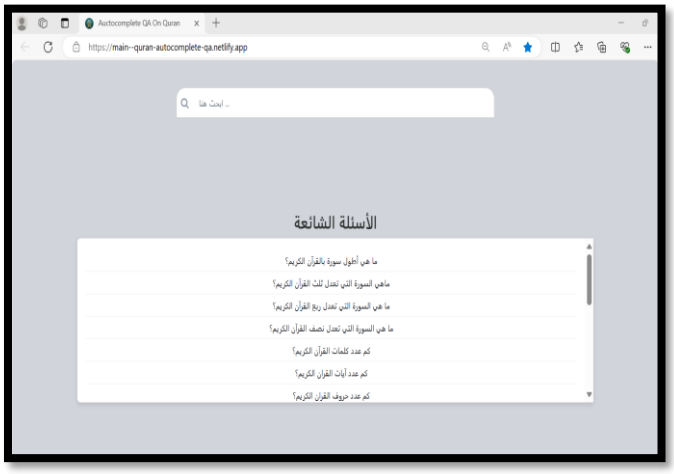

Figure 3 Web Interface of the Quran Autocomplete and Question Answering System.

G. Challenges and Solutions:

Handling the complexity of Arabic script and ensuring accurate similarity measurements between queries and stored data were major challenges. Solutions included implementing advanced text processing techniques and customizing similarity metrics to better suit the Arabic language's nuances.

V. Evaluation and Results

A. Evaluation Methodology:

The system's performance is evaluated using a blend of quantitative metrics and user experience assessments. We focus on several key areas:

1) ***Text Preprocessing and Normalization:*** The efficiency and accuracy of text normalization scripts are evaluated to ensure they are preparing the text appropriately for further NLP tasks.

2) ***Question-Answer Matching:*** We measure the effectiveness of our matching algorithms, particularly looking at how well the system identifies and ranks relevant answers to user queries.

3) ***Auto-completion and Spell Correction:*** The responsiveness and accuracy of the auto-completion and spell correction features are tested, noting how these enhance the user interaction.

B. Results:

The system's performance is illustrated through various screenshots showing the autocomplete and question-answering functionalities. These examples demonstrate the system's ability to handle a range of queries related to the Quran, providing accurate and contextually relevant answers.

Figure 4 Query for General Information about the Quran.

The pervious figure shows the system's response to a query about general information regarding the Quran. The autocomplete functionality offers suggestions as the user types, and upon submission, the system identifies the most relevant answer. The correct response is highlighted in the results, demonstrating the system's effective use of NLP to interpret and answer queries accurately.

The system effectively handles broad queries by providing both relevant autocomplete suggestions and accurate final answers, showcasing its robust knowledge base and semantic understanding.

Figure 5 Specific Query Regarding a Topic in the Quran.

In this example, the user inquiries about a specific topic within the Quran. The system not only suggests several closely related questions but also provides a precise answer to the chosen query, reflecting its capability to manage and retrieve detailed information.

This result highlights the system's strength in dealing with specific topics, facilitating users in exploring detailed aspects of the Quran through intuitive interactions. It underlines the precision of the NLP models in understanding and processing targeted inquiries.

Figure 6 Complex Query Involving Interpretive Information.

Here, the user asks a complex question about interpretive content in the Quran. The system's suggestions adapt as more details are typed, and it successfully delivers a comprehensive answer that addresses the query's intent.

This case illustrates the system's advanced capability to handle complex queries that require deep understanding and interpretation of the text. It demonstrates the effectiveness of the implemented NLP techniques, such as n-grams and Jaccard similarity, in providing contextually rich and accurate answers.

Figure 7 Query for Tafseer (interpretation) of a Specific Verse

This figure demonstrates the system's ability to respond to specific queries regarding the Tafseer (interpretation) of a verse. The user inputs a request for Tafseer, and the system accurately provides detailed explanations directly related to the queried verse.

The system's effectiveness in providing precise and detailed Tafseer highlights its utility for in-depth Quranic study. It not only confirms the system's capability to fetch specific information but also its ability to understand and process requests that require detailed textual analysis. This functionality is particularly valuable for scholars and students who require accurate interpretations of Quranic verses.

Figure 8 Query for Topics Related to Quranic Teachings.

The last case, the user searches for a general topic within the Quran. The system displays a list of related questions, allowing the user to explore various aspects of the topic. This feature facilitates a broader study of the Quran by linking related themes and inquiries.

This result underscores the system's robustness in handling broad thematic queries, demonstrating its comprehensive indexing and search capabilities. By providing a list of related questions, the system aids users in navigating complex religious texts, enriching their learning and exploration experience. The ability to connect different but related topics helps in building a more holistic understanding of the Quranic teachings.

VI. Conclusion and Future Work

A. *Conclusion:*

The Arabic Auto-complete Question Answering System for the Quran has demonstrated substantial capability in enhancing how users interact with and explore the Quranic text. By integrating advanced Natural Language Processing (NLP) techniques, such as text normalization, auto-completion, and similarity-based question answering, the system provides accurate and contextually relevant answers. This research successfully bridged the gap between traditional religious text study and modern computational methods, offering a tool that supports both educational and scholarly activities.

B. *Future Work:*

Looking forward, several enhancements and expansions are envisioned:

- ***Algorithmic Enhancements:*** Implementing deep learning techniques, like Recurrent Neural Networks (RNNs), could offer improvements in understanding the nuances of Quranic Arabic, potentially increasing the system's accuracy and response quality.

- ***Dataset Enrichment:*** Collaborating with Islamic scholars to augment the dataset with more diverse interpretations could enhance the system's comprehensiveness and accuracy. Including additional religious texts could also broaden the system's utility.

- ***Interface Improvements****:* Enhancing the user interface with additional features like voice recognition and support for other languages could make the system more accessible to a wider audience.

VII. Acknowledgment

The authors wish to express their deepest gratitude to Professor Adnan Yahya from the Department of Electrical and Computer Engineering at Birzeit University. His invaluable guidance and dedicated mentorship were instrumental in the successful completion of this project. We appreciate his extensive efforts and support throughout our research.

VIII. References

[1] G. M. B.-I. (. Marina Shorer-Zeltser, Applications of Content and NLP Techniques Analyses on Religious Internet Content, Common Ground Research Networks, 2008.

[2] L. Turner, Enhancing User Experience with Smart Auto-completion Systems, Stanford, 2018.

[3] M. R. R. M. A. M. B. F. N. E.-M. a. M. T. Heba Abdelnasser, Al-Bayan: An Arabic Question Answering System for the Holy Quran, Doha: Association for Computational Linguistics, 2014.

[4] "NLP project form," [Online]: https://www.jotform.com/form/241052004738043 [Accessed 20 6 2024].

[5] "Annotated Corpus of Arabic Al-Quran Question and Answer," [Online]: http://archive.researchdata.leeds.ac.uk/id/eprint/464 [Accessed 20 6 2024].