CSCE 5290: Natural Language Processing

Project Proposal

Title: Quran Memorization Assistant using NLP Technology

Group 1: Samir Tarda, Owais Jafer, Majed Alhazmi, Abdul Azeem Mohammed

1. Motivation

In the Islamic faith, the memorization of the Holy Quran is considered one of the most significant acts of worship, enabling Muslims to develop a profound connection with their creator, Allah. This practice not only serves as a means of preserving the sacred text but also represents a paramount spiritual aspiration for the vast majority, if not all, Muslims throughout their lives. However, the process of committing the Quran to memory poses several challenges for many individuals depending on their routines and lifestyle. Considering this, we propose to develop an NLP-based application that aims to streamline and facilitate the memorization of the Quran.

The primary objective of this project is to create a user-friendly app that can accurately predict the next word in a sequence, thereby assisting students in their memorization journey. The envisioned application would allow students to recite the Quran either verbally or through typing, and in the event of an error, the app would provide a series of warnings. Should the student fail to rectify their mistake, the app would offer the correct word or phrase, ensuring a smooth and accurate memorization process.

Additionally, if the project timeline permits, we will also incorporate advanced features that enable the app to comprehend the syntactic structures and contextual themes present in each sura (chapter) of the Quran. This functionality is meant to potentially allow users to customize personalized memorization plans tailored to their individual needs and preferences.

Through this innovative application, we seek to revolutionize the way Muslims approach Quranic memorization, making it more accessible, engaging, and rewarding for all those who desire to strengthen their bond with the divine word.

2. Significance

This project revolutionizes the traditional approach to Quran memorization by introducing a tool that blends spiritual practice with technological innovation. By providing real-time assistance and feedback, the app aims to mitigate common memorization challenges, thereby fostering a more inclusive and supportive learning environment. The significance of this endeavor lies in its potential to:

- **Enhance Learning Efficiency:** Tailored memorization plans and immediate error correction can significantly reduce the time and effort required to memorize the Quran.

- **Increase Accessibility:** Making memorization tools available on personal devices opens up opportunities for individuals with busy schedules or limited access to traditional learning resources.
- **Preserve Religious Tradition:** By facilitating easier memorization, the project contributes to the preservation and continuation of a vital Islamic practice.
- **Promote Spiritual Growth:** By removing barriers to memorization, more individuals can engage deeply with the Quran, fostering personal growth and a stronger connection with their faith.

3. Objectives

The project aims to achieve the following objectives:

- **Develop an NLP-based Application:** User-friendly app that can predict the next word in a Quranic verse, offering real-time assistance to users during their memorization process.
- **Incorporate Syntactic and Contextual Understanding:** The app comprehends the complex structures and themes of the Quran to provide contextually relevant assistance.
- **Tailored Memorization Plans:** Utilize user data and preferences to generate personalized memorization strategies that align with individual learning styles and goals.
- **Measure Success:** Implement a system for tracking progress and feedback to continually refine the app's effectiveness and user satisfaction.

4. Features

- **Real-Time Word Prediction and Error Correction:** Utilizing advanced NLP algorithms to offer immediate assistance and correction, enhancing memorization accuracy.
- **Contextual Awareness:** Analysis of verses to understand themes and structures, ensuring that assistance is both relevant and respectful of the text's sanctity.
- **Personalized Learning Plans:** Algorithms that assess user performance and preferences to adapt the learning process accordingly.
- **User Progress Tracking:** Features to monitor memorization progress, offering insights and motivation to users.

5. Dataset.

The project will utilize the Quranic Arabic Corpus, which includes detailed morphological, syntactic, and semantic annotations of the Holy Quran. This rich dataset not only provides the Arabic text but also includes information on word-by-word analysis, grammatical tagging, and root derivations. This comprehensive data is invaluable for developing an NLP application aimed at assisting in the memorization of the Quran, as it allows for deep linguistic and contextual analysis.

5.1 Dataset Size and Type: The Quranic Arabic Corpus covers the entire text of the Quran, consisting of 114 chapters (Surahs), over 6,000 verses (Ayahs), and tens of thousands of words. The dataset includes morphological tagging, syntactic annotations, and the lemma (base form) of each word. The annotations also detail the grammatical features of words, such as part of speech, gender, number, and case.

5.2 Dataset Source: The primary source of the Arabic text is the Tanzil project, renowned for its accuracy and verification process. The annotations have been provided by the Quranic Arabic Corpus project, ensuring respect for copyright and terms of use.

5.3 Dataset Preprocessing: To ensure the dataset's suitability for NLP analysis and application development, the following preprocessing steps will be implemented:

- **Data Cleaning:** Verify the integrity of the text and annotations, ensuring there are no missing or duplicated entries.
- **Normalization:** Standardize the representation of the Arabic text, including the harmonization of orthographic variants and the removal of diacritical marks that are not essential for the project.
- **Tokenization:** Break down the text into manageable units (words and verses) for easier processing and analysis.
- **Lemmatization:** Utilize the lemma annotations to group together the inflected forms of a word so they can be analyzed as a single item.
- **Feature Extraction:** Extract and structure the morphological, syntactic, and semantic features from the annotated data for use in the application's algorithms.

5.4 Dataset Utilization and Implementation: Our application will be designed to not only aid in memorization but also to enrich the user's understanding of the Quranic text, fostering a deeper connection with the sacred scripture. The dataset will be employed to:

- Train the NLP model to understand and predict Quranic Arabic text accurately.
- Enable word-by-word assistance in memorization, providing not just the next words but also their meanings and grammatical context.
- Develop personalized memorization plans based on the syntactic and thematic structure of the verses. This is an advanced feature and will be integrated depending on the timeline of the project.

6. Visualization

Our project's structure, data flow, and overall strategy, we will develop a series of visualizations that highlight the key components and processes involved in the development of the application. These visualizations demonstrate the project’s technical details and broader goals.

6.1 System Architecture Diagram: This high-level diagram outlines the core components of the application, including:

- **User Interface (UI):** Displays the front-end components where users would interact with the app, such as the text input field or voice input interface, and the app's feedback mechanisms.
- **NLP Engine:** represents the central processing unit of the application, including text processing, prediction algorithms, and error detection.
- **Data Storage:** Show the database where user data, progress, and preferences are stored for personalized memorization plans.
- **External Resources:** Highlight the integration with the Quranic Arabic Corpus for accessing the text and annotations.

6.2. Data Flow Diagram: A detailed data flow diagram (DFD) will depict how data moves through the system, from user input to the final output. Key processes illustrated will include:

- **Input Processing:** How the app handles and analyzes typed or spoken input from the user.
- **Word Prediction & Correction:** The process of predicting the next word and providing corrections, including how the NLP engine accesses linguistic data.
- **Personalization Algorithm:** The flow of data that tailors the memorization plan to the user's performance and preferences.

6.3 User Interaction Flowchart: focuses on the user's journey through the application, showcasing:

- **Starting a Session:** Steps from launching the app to beginning a memorization session.
- **During a Session:** User actions and app responses, including correct predictions, error detections, and corrections.
- **Review and Progress Tracking:** How users can review their progress and adjust their memorization plans.

6.4 NLP Model Training Process: A diagram illustrating the stages involved in training the NLP model with the Quranic Arabic Corpus, highlighting:

- **Preprocessing:** demonstrates the steps taken to clean and prepare the dataset.
- **Feature Extraction:** involves the extraction of relevant features for analysis.
- **Model Training**: illustrates the training cycle, including validation and testing phases.

6.5 Personalized Memorization Plan Algorithm: A schematic representation of how the app generates personalized memorization plans, showing:

- **User Data Analysis:** Evaluation of user's performance to identify strengths and weaknesses.
- **Plan Generation:** Algorithmic selection of verses and sections for focused memorization based on analysis.
- **Adaptation Over Time:** this involves capabilities to adjust plans based on ongoing user performance and feedback.

Note: we ran out of time to create the actual visual diagrams but we can provide them in our next submission should we need to.

7. Dataset Resources & References:

1. https://corpus.quran.com/
2. https://corpus.quran.com/download/
3. https://github.com/quran
4. https://corpus.quran.com/java/api/index.html
5. https://corpus.quran.com/java/api/overview-summary.html

8. Github: https://github.com/sntarda/nlp