# A Multimodal Data Scraping Tool for Collecting Authentic Islamic Text Datasets

Abdallah Namoun[1], Mohammad Ali Humayun[2], Waqas Nawaz[3]
AI Center, Faculty of Computer and Information Systems, Islamic University of Madinah, Madinah, 42351, Saudi Arabia[1, 3]
Department of Artificial Intelligence, Information Technology University, Lahore, 54600, Pakistan[2]

***Abstract*—Making decisions based on accurate knowledge is agreed upon to provide ample opportunities in different walks of life. Machine learning and natural language processing (NLP) systems, such as Large Language Models, may use unrecognized sources of Islamic content to fuel their predictive models, which could often lead to incorrect judgments and rulings. This article presents the development of an automated method with four distinct algorithms for text extraction from static websites, dynamic websites, YouTube videos with transcripts, and for speech-to-text conversion from videos without transcripts, particularly targeting Islamic knowledge text. The tool is tested by collecting a reliable Islamic knowledge dataset from authentic sources in Saudi Arabia. We scraped Islamic content in Arabic from text websites of prominent scholars and YouTube channels administered by five authorized agencies in Saudi Arabia. These agencies include the general authority for the affairs of the grand mosque and the prophet's mosque and charitable foundations in Saudi Arabia. For websites, text data were scraped using Python tools for static and dynamic web scraping such as Beautiful Soup and Selenium. For YouTube channels, data were scraped from existing transcripts or transcribed using automatic speech recognition tools. The final Islamic content dataset comprises 31225 records from regulated sources. Our Islamic knowledge dataset can be used to develop accurate Islamic question answering, AI chatbots and other NLP systems.**

***Keywords*—*Web scraping; Islamic knowledge; machine learning; natural language processing; question and answering; AI chatbots***

## I. Introduction

We propose a multimodal web scraping tool to collect Islamic content from verified websites and YouTube channels administered by five trusted entities in Saudi Arabia. Our crawling algorithms use Python libraries, including Beautiful Soup and Selenium, to scrape content from static and dynamic websites and channels. We selected input sources from official websites that are administered and maintained by non-profit organizations (e.g., agencies and charities) in Saudi Arabia. This strategy is adopted to achieve two important objectives: first, to ensure the credibility and correctness of the Islamic data published on these sources, and second, to eliminate noise data from untrusted sources, which could sabotage the quality of our proposed dataset.

Next, we exemplify the application of our tool on Islamic data from authentic Islamic knowledge websites that are maintained by 1) authorized non-profit organizations represented by the General Authority for the affairs of the Grand Mosque and the Prophet's Mosque (also known as Alharamain) or 2) under the management of charitable organizations in Saudi Arabia for scholars who are past or present members of the Permanent Committee for Scholarly Research and Ifta and/or Saudi Council of Senior Scholars.

To contextualize our research problem, let us consider the subsequent realistic scenario, which describes the typical challenges faced by non-Arabic-speaking people when searching for Islamic knowledge and teaching to guide their lifestyle decisions. Decisions may sometimes impact a person's life, as outlined below.

"Ahmed is a 35-year-old married government employee who has recently faced a family hardship concerning the decision to abort his child due to health concerns regarding his wife. He comes from a small village in Pakistan, and he does not speak the Arabic language. However, as a devoted Muslim, Ahmed is never satisfied with simple answers and always strives to make knowledge-driven decisions in search of satisfaction and tranquility. Ahmed starts by asking his village's main scholar at the nearest mosque; however, he remains unconvinced about the answers he received. In his quest for more resounding answers, he stumbles across a QA NLP platform, which offers opinion-oriented knowledge translated into various languages. Luckily, Ahmed can query the online platform on his mobile device, using plain text in Urdu, about the topic of child abortion in Islam under different circumstances and scenarios. The language-driven system presents the answers in a concise form with pieces of evidence from different scholars by extrapolating its unique and authentic dataset of Islamic knowledge. Not only can Ahmed read the answers in Urdu, but he can also view different opinions accompanied by arguments linked to the sources of the Quran and Sunnah. Ahmed can now discuss these options openly with his family before choosing the best option given his family's circumstances."

Reading this simple hypothetical scenario, various requirements emerge that should be addressed to effectively interpret Islamic knowledge.

- The non-Arabic speaking community constantly needs to access information using modern technologies, such as mobile devices.
- Intelligent NLP systems require reliable Islamic data to be able to produce meaningful and trustworthy knowledge.
- Our literature review (in the next section) revealed the lack of datasets that serve the objectives of the above scenario. Moreover, there are other scientific motivations

This research is supported by the Deanship of Scientific Research of the Islamic University of Madinah, KSA under the research groups (first) project no. 956.

and advantages to creating our unique web content collection tool and Islamic dataset, as elaborated below.

- We developed an automated data collection tool that gathers and organizes multimodal data from publicly available, reliable Islamic text and audio-video sources on the web, with minimal manual intervention, and compiles it into a text dataset. Fellow researchers can customize our tool to scrape similar content from online sources.
- We provide an authentic dataset of Islamic knowledge extracted from approved and regulated online sources. This uniqueness distinguishes our dataset from published Islamic content datasets. The proposed dataset can be used to form the foundation for additional authentic datasets and extend existing Islamic ruling datasets. Researchers from different specializations (e.g., Islamic studies, theology, sociology, and history) may use the datasets to conduct research studies and perform rigorous analyses.
- Our Islamic knowledge dataset may be reused by fellow machine learning researchers for 1) developing and testing machine learning models of generative AI systems for understanding and issuing Islamic rulings, 2) extracting the reasons behind certain rulings and judgments, 3) comparing rulings among various schools of thoughts and regions to understand their commonalities and differences, 4) exploring the evolution of Islamic laws and rulings over time in response to the contemporary technological and social changes, and 5) investigating the current issues in Muslim communities. Examples of natural language processing (NLP) applications on Islamic content and text include machine translation, question answering, Islamic rulings classification, text summarization, text analytics, auto-diacritization, and smart assistants and AI chatbots.
- Providing authentic Islamic datasets from trusted sources can assist policymakers, local organizations and centers, and the private sector to revise and/or create their policies around the Islamic principles within the datasets. For instance, financial products and services in banks may be offered using policies extracted from this Islamic knowledge. Similarly, the food industry may utilize the certifications and regulations pertaining to halal products from the knowledge extracted from these datasets. It may also guide courts and legal practitioners in Muslim countries to make informed decisions and rulings.
- Digitizing and archiving authentic datasets of Islamic knowledge, including rare texts and manuscripts, helps preserve the Islamic culture and heritage for the next generations. Moreover, these datasets can be translated into other languages to benefit non-Arabic audiences and communities.
- This Islamic dataset can be used to promote interfaith dialogue by providing accurate and authentic information about Islamic beliefs, practices, and rulings, especially concerning contemporary issues. It creates an opportunity to enhance public knowledge about Islam, reducing misconceptions and prejudice.

## II. Related Work

Collecting and compiling authentic and reliable datasets of Islamic content and knowledge is crucial for training and validating natural language processing (NLP) and generative artificial intelligence (GenAI) models [1] which can be subsequently consumed by ordinary users and local entities (e.g., courts). However, the Internet offers various unverified sources of Islamic content hosted on several websites and social media. It is challenging to judge the accuracy and relevance of such Islamic knowledge to train relevant generative artificial intelligence models.

Indeed, there is a lack of trustworthy Islamic datasets that discuss essential topics and perspectives on societal issues concerning Muslim communities. A recent survey compared 11 question-and-answering (Q&A) Islamic datasets published between 2014 and 2022 [2]. All these corpora covered Quran questions, overlooking other important sources of Islamic knowledge and rulings. The authors in study [3] introduced an Islamic dataset (i.e., eiad) for building English question-answer AI chatbots. The dataset is in English and covers 15 categories, targeting converts and non-Muslims. The dataset comprises 10000 articles collected from three major, trusted websites like IslamQA.com. Similarly, the authors in study [4] contribute a dataset (i.e., QASiNa) in the Indonesian language for question-answering tasks. This dataset is unique since it is based on evidence from the Sirah literature (i.e., the Prophet Muhammed practices and sayings). Such datasets are useful for training large language models, such as ChatGPT and Gemini.

The study published recently in [5] is one of the exceptions. It is claimed to be the first Islamic rulings (i.e., is called fatwaset in Arabic) dataset. The data were collected from 13 trusted websites in Arab countries, including Saudi Arabia, Egypt, Jordan, Qatar, and Syria. The authors performed an exploratory data analysis revealing a total of 130182 records. Authors in study [1] suggest a dataset for developing chatbot systems for the issuance of Islamic fatwas. The dataset is claimed to be the largest Fatwas dataset, with the classification of topics. In total, the dataset has approximately 850000 fatwas, scraped from different websites, regions, and schools of thought. It is observed that most Fatwas (71%) were extracted from AskFM.

Authors in study [6] proposed an Arabic Multi-IsnadSet (MIS) dataset as a Neo4j multi-directed graph comprising 2029 narrator nodes and 77797 sanad-hadith connections. Hadith include the sayings, actions, or silent approvals of the Prophet Muhammad. Sanad represents the credibility (through a chain of narrators) of each hadith in the dataset. Hadith is the second source of religious legislation in Islam, after Quran. Data scraping tools were used to fetch hadith details, such as hadith number, content, list and sequence of narrators, and Isnad count. This dataset allows researchers to understand the authenticity of each hadith and the strength of its narration. Similarly, authors in [7] compiled a dataset of 650K hadiths, named Sanadset. The proposed dataset collected from 926 Arabic books may be used for automatic sanad verification and classification. However, the dataset does not distinguish between the authenticity of each hadith.

Although the existing works have collected diverse datasets across multiple domains, a unified dataset comprising data from multimodal sources, such as video lectures and text sources of different nature is still lacking. Besides, most of the works have relied on manual compilation and organization of datasets, disregarding automation strategies for dataset scraping and compilation. Furthermore, insufficient effort has been made to ensure that the sources are authentic and verified by an authoritative body. These limitations affect the ability of these datasets to adapt to future needs.

This work aims to fill this gap by designing an automation strategy for dataset collection. Moreover, the strategy incorporates datasets from multimodal sources, such as video and text resources, ensuring that all sources are authentic. This approach not only makes the dataset both authentic and diverse but also facilitates its easy expansion and adaptation for multiple domains.

## III. Materials and Methods

Our web sources from which we curated our dataset are authenticated and administered by organizations and foundations in Saudi Arabia are listed in Table I. Our dataset is publicly accessible at [8]. Fig. 1 shows samples of Youtube channels.

TABLE I. Target Reliable Sources of Islamic Knowledge

| Name | Administration | URL Links |
|---|---|---|
| Presidency of the Two Holy Mosques | General Authority for the affairs of the Grand Mosque and the Prophet's Mosque | gph.gov.sa/index.php/ar/ |
| | | manaratalharamain.gov.sa/speeches/ |
| | | youtube.com/@twjehDM |
| | | youtube.com/@makkah |
| | | youtube.com/@SaudiQuranTv |
| Presidency of the Two Holy Mosques (Manarat Al Haramain) | Agency of the Affairs of Al-Masjid Al-Nabawi | wmn.gov.sa/public/ |
| | | youtube.com/@wmngovksa |
| | | youtube.com/@wmngovsa |
| | | youtube.com/@SaudiSunnahTv |
| Ibn Baz | Sheikh Abdul Aziz bin Baz Charitable Foundation | binbaz.org.sa/ |
| Ibn Othaimeen | Sheikh Mohammed bin Saleh Al Othaimeen Charitable Foundation | binothaimeen.net/ |
| | | youtube.com/channel/UCtF3YygTiodnYSw8vD3UJtQ |
| Alfawzan | Al Dawa Charitable Foundation | alfawzan.af.org.sa/ar |
| | | youtube.com/@salihalfawzan |
| | | youtube.com/@aforgsa1 |

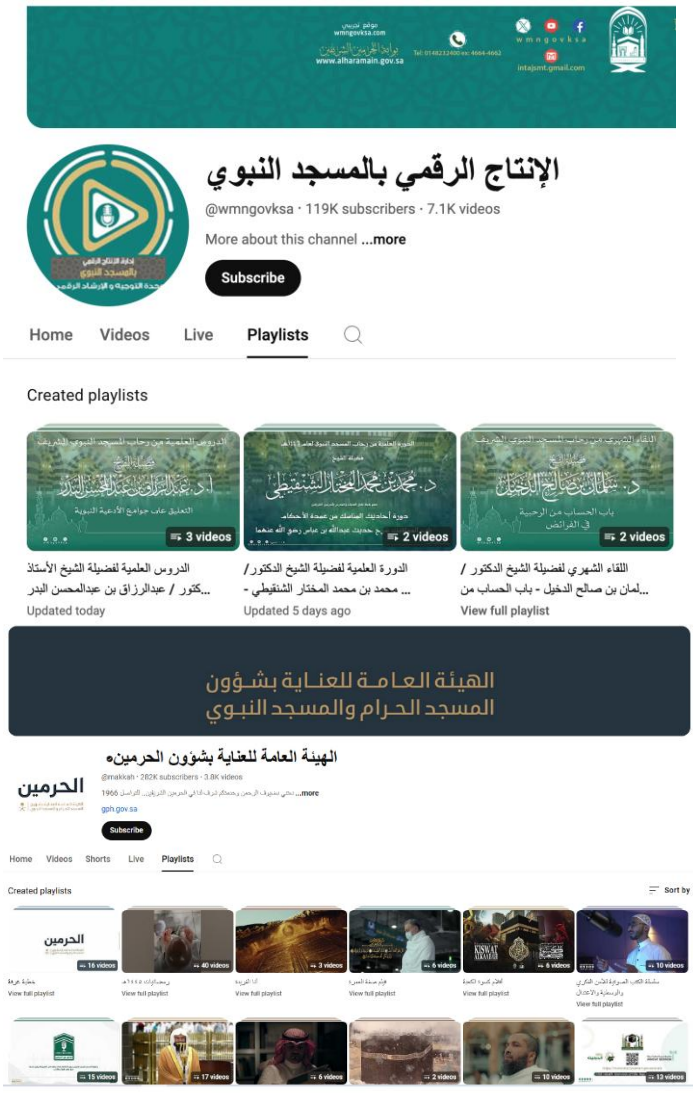
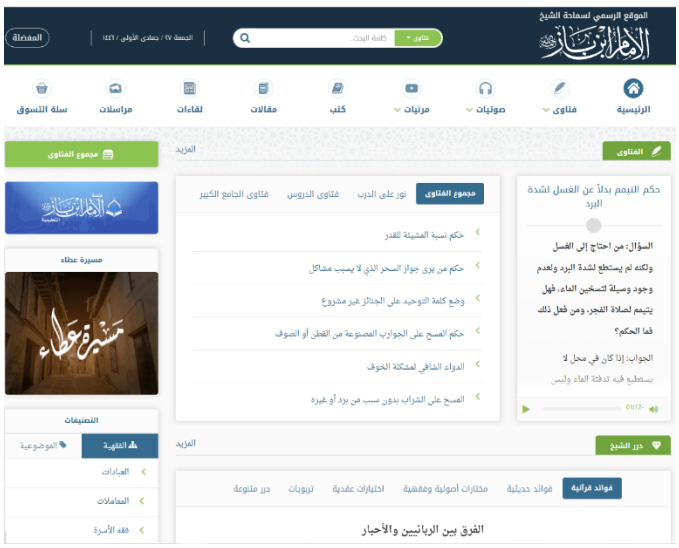



Fig. 1. A sample of YouTube channels.



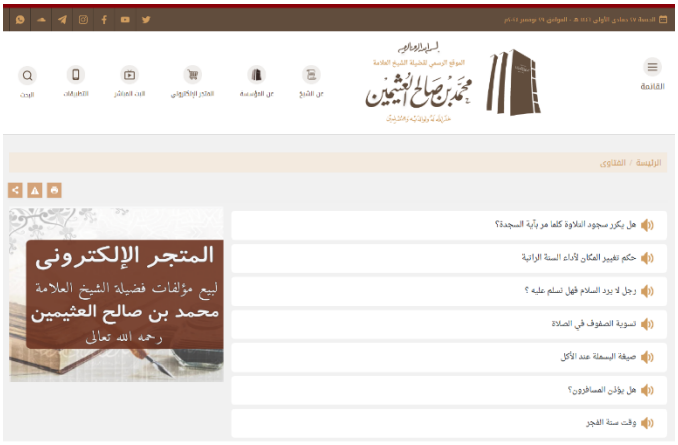



Fig. 2. A sample of Islamic websites.

TABLE II. TYPE OF DOWNLOADABLE DATA FROM NON-PROFIT ORGANIZATIONS

| Data Category | Bin Othaimeen (Mix) | Bin Baz (Mainly Text) | Alfawzan (Audio) |
|---|---|---|---|
| Articles المقالات | No | Yes | Yes |
| Lessons الدروس | Yes | Yes | Yes |
| Lectures المحاضرات | Yes | No | Yes |
| Meetings and Dialogues اللقاءات والحوارات | Yes | Yes | No |
| Rulings الفتاوى | Yes | Yes | Yes |
| Speech الكلمات | No | No | No |
| Letters - Communications الرسائل – المراسلات– الخطابات | No | Yes | No |
| Sermons (not on Fridays) الخطب | Yes | No | Yes |
| Friday Sermons خطب الجمعة | No | No | No |

We applied the following inclusion and exclusion criteria (see Table III) to identify the sources of our Islamic dataset. When it was not possible to retrieve the text or transcripts, we used speech-to-text (Arabic audio transcription) API services to transcribe the videos of Islamic knowledge. Table II shows type of downloadable data from non-profit organizations.

## IV. SOFTWARE ALGORITHMS

Mainly, four different algorithms have been developed: one for extracting text data from static websites, another for retrieving text data from dynamic websites, a third for obtaining relevant transcripts from YouTube videos with available transcripts, and the last for translating speech-to-text from videos without a transcript. The code used is publicly available at [9] along with usage instructions, and the dataset has been uploaded to [9].

TABLE III. INCLUSION AND EXCLUSION CRITERIA

| Criterion | Inclusion | Exclusion |
|---|---|---|
| Sources | Authentic sources that are maintained by local authorities | Unknown sources or not managed by authorized entities |
| Responsible | Organization, foundation, or charity | Individual |
| Purpose | Non-profit | Commercial |
| Country | Saudi Arabia | Other countries |
| Type of data | Opinions about different topics published in various form (e.g., speeches, letters, sermons, lectures, explanations, etc.) | Books, booklets, and handwritten manuscripts |
| Format | Downloadable text, audio, video files (with ability to scrape data) | Images |
| Language | Arabic | Other languages (English, Urdu, … etc) |
| Place of Publication | Knowledge published on official websites and YouTube channels | Knowledge posted on social media platforms, such as X (Twitter) platform and Facebook |

### A. Text Sources

Only the websites owned by reputable scholars from Saudi Arabia were selected for the text data. Consequently, the official websites of Ibn Baz, Ibn Othaimeen, and Alfawzan were chosen to fetch data. Since the websites have different structures and data types, the web scraping methodology [9] for fetching data from each website differed.

The website for Ibn Baz employs static HTML content, so the Python Beautiful Soup library was used to scrape data from this site by exploiting the HTML tags. The website has different categories of data including: الفتاوى (Islamic rulings), مقالات (articles), لقاءات وحوارات (discussions), خطابات ومراسل (speeches), and درر (pearls). First, all fatwas were fetched; these fatwas are subdivided into four major categories for each radio program نور فتاوى الجامع الكبير, and على الدرب , فتاوى لدروس, مجموع الفتاوى Each of these categories of fatwas has multiple pages. To extract data from each page, the base URL for the fatwa page was used, and then iterations were performed over the different program categories to construct the URLs for each program. The total number of pages varied for each fatwa category, and URLs were appended by iterating through the pages until the last page for each program was reached and scraped. Fig. 2 shows samples of Islamic websites.

The HTML data from the final URL for a page within the category was fetched using the Python requests library for HTTP requests. An HTTP request was sent to the URL, and the HTML response was filtered using Beautiful Soup to fetch the URLs for subpages pointing to the individual fatwas. The pages did not contain only URLs for the fatwa but also other URLs, such as those pointing to the homepage and other categories. Therefore,

only the URLs containing the keyword corresponding to the data category were filtered out. For example, URLs with fatwa had the keyword "fatwa" in them, so URLs were retained accordingly from each fatwa.

Subsequently, the filtered URLs were iterated, and HTTP requests were sent to each page to fetch the HTML response, which was again cleaned using Beautiful Soup. The HTML response from each fatwa page was analyzed using Beautiful Soup, and the contents for the following fatwa fields were fetched: title, question, answer, and category. HTML tags for each field were looked up, and the corresponding text was extracted. The fetched texts were stored in the dataset within different fields along with the main category (fatwa), the fatwa program, the page number, and the final URL for the fatwa.

For other text categories, such as مقالات (articles), لقاءات وحوارات (discussions), خطابات ومراسل (speeches), and درر (pearls), a similar process was applied to fetch the data with a few changes. These categories do not have different radio programs, so a single loop iterates over the page numbers within the main URL. After the final content URLs were scraped using Beautiful Soup and by filtering the correct keywords of each of these categories, the content URL was fetched. Pages of these categories did not have fields like fatwa category and question, so only title and text fields were compiled.

The dataset was cleaned to remove missing entries and duplicate entries added from multiple site URLs. Finally, due to different field structures, the dataset is stored as two separate CSV files: one for fatwas and one for other miscellaneous data.

The website for Alfawzan had similar static content, so the methodology was almost the same as the one for Ibn Baz website, with a few changes. Firstly, all fatwas on Alfawzan website have not been subdivided according to different programs but are placed under different pages in the main fatwa URL. Hence, the loop for adding page numbers was run with the main fatwa URL only. Secondly, most of the fatwas on the website have audio content and do not have text transcriptions, as indicated by a text file button being greyed out for fatwa without text content. Consequently, before fetching the reference of the fatwa URL and sending an HTTP request, a check was performed to see if the text button was not greyed out. Only then was the subsequent HTTP request sent to the URL, and the response HTML content filtered for the fatwa text fields like category, title, question, and answer.

Another source of text data is the website by Ibn Othaimeen. This website has a rich collection of fatwas in different programs, but the challenge was that it does not store static data. Instead, the site uses JavaScript to load its content dynamically. Hence, HTML-based filtering and the Beautiful Soup library could not be implemented for this site.

We used the Python Selenium library for this website to simulate browsing using the Chrome driver. This allowed us to fetch data by simulating browsing, scrolling, clicking, and recording the response. We used Selenium to control the Chrome driver in headless mode so that the GUI is not displayed, but the operations are performed seamlessly in the background.

The fatwas on this site are organized mainly into three different programs, namely: فتاوى نور على الدرب, لقاءات الباب المفتوح and اللقاء الشهري. All three have different URLs and different numbers of pages under the main URL, which contain links to the individual episodes. These episodes, in turn, had links to the fatwa. The episode numbers for فتاوى نور على الدرب are named as الشريط, while for the other two programs they are named the same as the program with numbering. So, the hierarchy of data was: fatwa > radio program > episode number > fatwa.

First, we iterated a loop over the URLs for the three programs. For each program, we further iterated over all the pages to append to the base URL for the program. Then, we used the Selenium Python library to make the Chrome driver go to the program URL with the page number and get the response to fetch all the web elements on the page.

As the page had multiple web elements, we were interested in only those pointing to the subsequent episode links containing the fatwa, e.g., those named with رقم الشريط for فتاوى نور على الدرب and so on. Hence, we filtered web elements with the corresponding text in their XML path. Then, we iterated over all those filtered elements on the page, got the text attribute from each element, and stored them in a list.

After that, we iterated over the list of texts and fetched the corresponding web elements to execute the JavaScript code to scroll to that web element and then execute the click for the element. After clicking the element, we waited until the next page was loaded. Once the episode page loaded and the web driver was pointing to the episode page, we got the corresponding URL of the episode page and stored it in the list. Similarly, we went back and got the URL for all the web elements on the episode page.

Eventually, we had the URL for all the episodes on a particular page within a certain radio program. By iterating over all the pages and all the programs, we got the URL for all episodes. Then, we iterated over the obtained episode URLs and set the Selenium-controlled driver to go to each URL iteratively.

Once all fatwa links were loaded and clickable on the episode page, we filtered the fatwa-referencing web elements by the condition of the XML path style containing a cursor pointer. We waited for all these elements to be fully loaded and clickable, then got texts for each.

Finally, the elements referencing each fatwa were fetched by looking for the texts obtained from their icons in the previous step. The corresponding web elements were clicked one by one. On clicking the fatwa element, the Chrome driver finally landed on the page containing the fatwa content, which included the title of the fatwa, the question, the answer, and its hierarchy path on the website (see Fig. 3).

When the fatwa web page loaded completely, the three different elements were fetched one by one, i.e., for the title, question, and answer, based on the CSS selectors 'p.title', 'div.fatwah-ques-cont', and 'div.fatwah-ans-cont' respectively. For each of these web elements, the inner HTML attribute was fetched, which loaded the HTML content. Finally, the Beautiful Soup library was used to fetch the required text from the fetched HTML for the element. This text data was stored in a CSV file under the following fields for each fatwa: title, question, answer, current-url, and category.

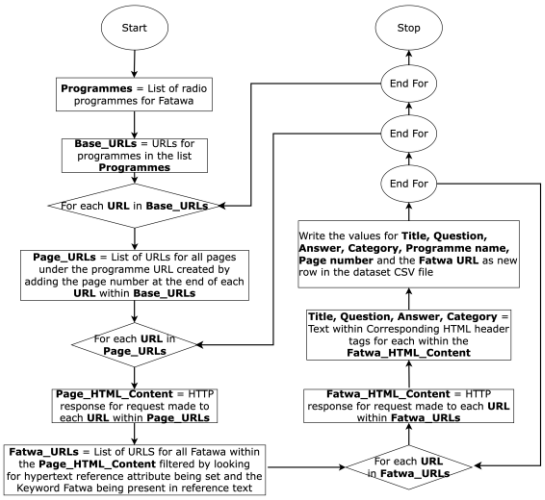



Fig. 3. The flowchart depicting the logic for collecting fatwa text from static websites.

## B. Youtube Transcripts

To get the transcripts from YouTube videos, official YouTube channels approved by the respective organizations were targeted, and a list of playlists to be targeted was identified for each channel. For each playlist, a list of URLs for all videos within that playlist was fetched using the playlist method from the Pytube Python library. Subsequently, each video URL within the list was looped through to get the transcript for the video using the get transcript method of the YouTube transcript API in Python. In case a video didn't have a transcript on YouTube, it returned a "transcript not found" error, in which case the URL of the video was added to a list of missed videos for that channel to be targeted for speech recognition.

Transcripts fetched for videos had text transcripts and timestamps, which were retained. Another clean copy of the transcript was also created, removing the timestamps and retaining the complete script as a single entity. Both the timestamped and the clean transcripts, along with the video title and URL, were saved in a separate SQL database file for that channel (see Fig. 4 and Fig. 5).

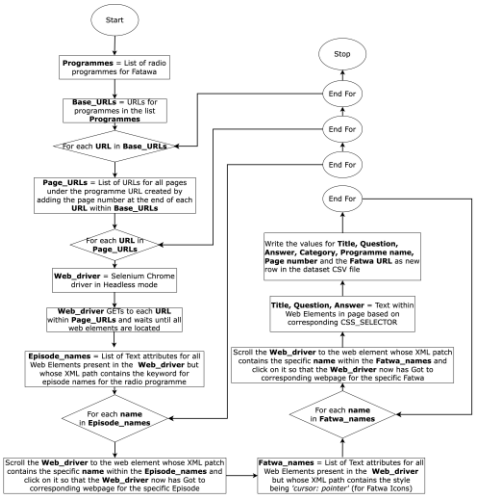



Fig. 4. The flowchart for collecting content from a website with dynamic data.

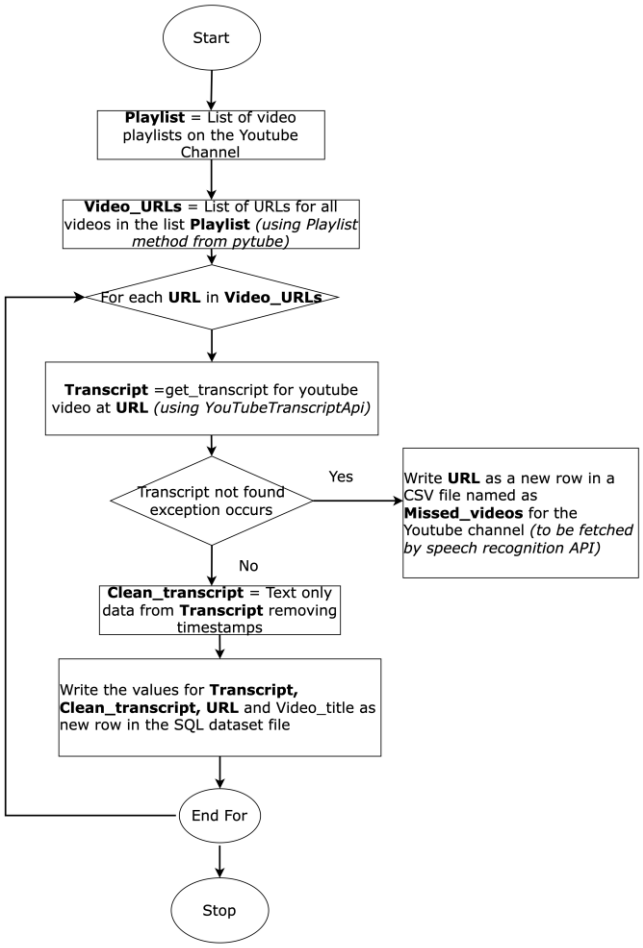



Fig. 5. The flowchart for collecting transcripts from YouTube videos.

## C. Automatic Speech Recognition

The file of missed videos created while getting the transcripts for each channel was then used to get the scripts of these videos with Automatic Speech Recognition (ASR) using the Google Speech Recognition API in Python. First, for each URL in the list of missed video URLs, the corresponding video was downloaded using the streams method of the YouTube class in the pytube library. The video was downloaded in MP4 format, and its audio was extracted and saved as a WAV audio file using the audio method of the VideoFileClip class from the moviepy.editor Python library. The converted WAV audio file was then used for speech recognition. However, longer files often have sections that are not recognized by the speech recognition engine, causing errors. To address this issue, the audio file was segmented into smaller chunks of approximately 180 seconds each and passed to the Arabic speech recognition engine of the Google Speech Recognition API. The text for each chunk was recognized and stored in a text file. Once all the segments for a video were processed, the text chunks were concatenated and saved as a single transcript for the video, along with the video URL and title, in an SQL DB file. This process generated the ASR scripts for the channel (see Fig. 6).

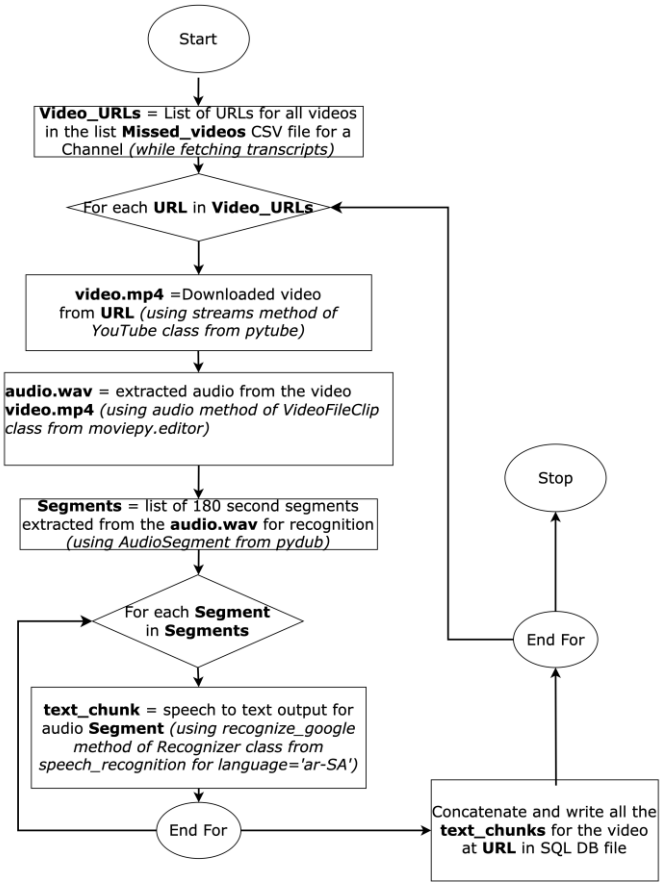



Fig. 6. The flowchart for collecting ASR text from YouTube videos.

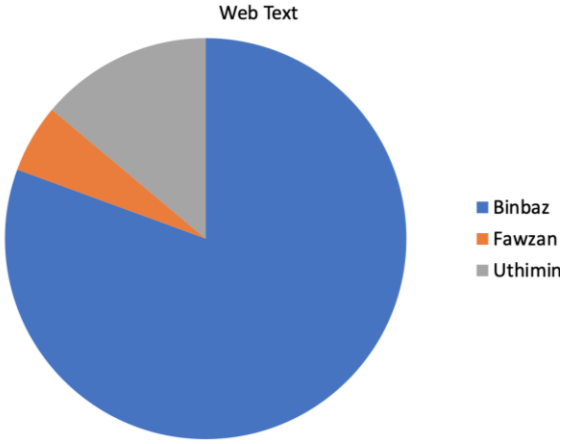
## V. Dataset Statistics

The resulting Islamic content dataset is organized into three main folders, namely Youtube_transcripts, Youtube_asr, and Text_data (see Table IV). The Youtube Transcripts folder contains seven files with 1363 records. The Youtube ASR folder contains eight files with 2972 records. Text data folder contains four files with 31225 records.

TABLE IV. Islamic Data Files and their Characteristics in our Dataset [A Total of 35560 Records]

| Data source | File Name | Records | Storage format | Owner |
|---|---|---|---|---|
| Youtube Transcripts | Makkah_transcripts.db | 242 | SQL DB | General Authority for the affairs of the Grand Mosque and the Prophet's Mosque |
| Youtube Transcripts | SaudiQuranTv_transcripts.db | 75 | SQL DB | Saudi Broadcasting Authority |
| Youtube Transcripts | wmngovsa_transcripts.db | 9 | SQL DB | Agency of General Presidency for the Affairs of Al-Masjid Al-Nabawi |
| Youtube Transcripts | SaudiSunnahTv_transcripts.db | 281 | SQL DB | Radio and Television Corporation in the Kingdom of Saudi Arabia |
| Youtube Transcripts | wmngovksa_transcripts.db | 263 | SQL DB | Agency of General Presidency for the Affairs of Al-Masjid Al-Nabawi |
| Youtube Transcripts | twjehDM_transcripts.db | 334 | SQL DB | General Authority for the affairs of the Grand Mosque and the Prophet's Mosque |
| Youtube Transcripts | ibnothaimeentv_transcripts.db | 159 | SQL DB | Sheikh Muhammad bin Saleh Foundation |
| youtube_ASR | Makkah_asr.db | 421 | SQL DB | General Authority for the affairs of the Grand Mosque and the Prophet's Mosque |
| youtube_ASR | SaudiQuranTv_asr.db | 9 | SQL DB | Saudi Broadcasting Authority |
| youtube_ASR | SaudiSunnahTv_asr.db | 126 | SQL DB | Radio and Television Corporation in the Kingdom of Saudi Arabia |
| youtube_ASR | salihalfawzan_asr.db | 1016 | SQL DB | Al Dawa Charitable Foundation |
| youtube_ASR | aforgsa1_asr.db | 92 | SQL DB | Al Dawa Charitable Foundation |
| youtube_ASR | twjehDM_asr.db | 218 | SQL DB | General Authority for the affairs of the Grand Mosque and the Prophet's Mosque |
| youtube_ASR | ibnothaimeentv_asr.db | 614 | SQL DB | Sheikh Muhammad bin Saleh Foundation |
| youtube_ASR | wmngovksa_asr.db | 476 | SQL DB | Agency of General Presidency for the Affairs of Al-Masjid Al-Nabawi |
| Web text | binbaz_misc_final_v1.csv | 718 | CSV | Sheikh Abdul Aziz bin Baz Charitable Foundation |
| Web text | binbaz_fatwa_final_v1.csv | 24448 | CSV | Sheikh Abdul Aziz bin Baz Charitable Foundation |
| Web text | othaimeen_fatwa_v1.csv | 4334 | CSV | Sheikh Muhammad bin Saleh Foundation |
| Web text | alfawzan_fatwa_v1.csv | 1725 | CSV | Al Dawa Charitable Foundation |

Fig. 7 shows the proportion of records in the final dataset from different sources.



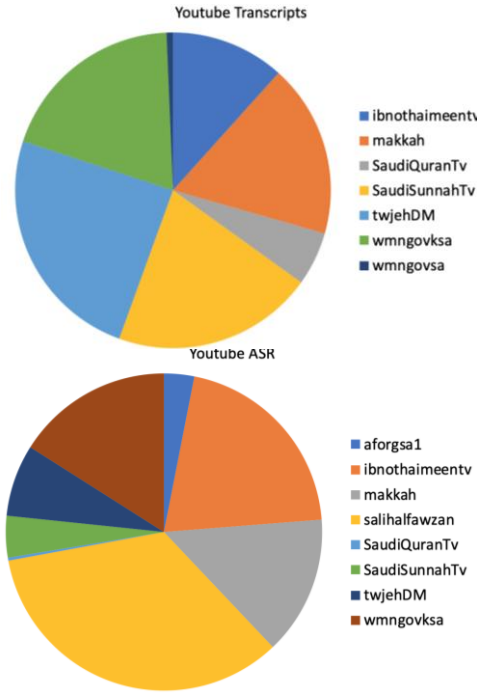



Fig. 7. Total records in the final dataset from different sources.

The focus of the example implementation has been on compiling authentic Islamic datasets, showcasing the capability of the algorithm to collect specialized datasets. However, the approach is generalizable and can be used in multilingual and multi-domain tasks, offering a versatile tool for dataset collection in various fields.

The code has been organized and publicized on a GitHub repository [9]. The repository contains different modules for scraping static and dynamic websites, as well as for fetching transcripts and performing speech-to-text conversions from YouTube videos. To use the code, users need to run the scraping.py file, providing the corresponding options for input and output files, and base URLs for the sites to fetch data from. Additionally, an action option must be specified to indicate whether to use dynamic, static, transcript, or ASR functionality. This will call the appropriate function within the scraping.py file. Fig. 8 is a screenshot from Github repo indicating the structure of code files.

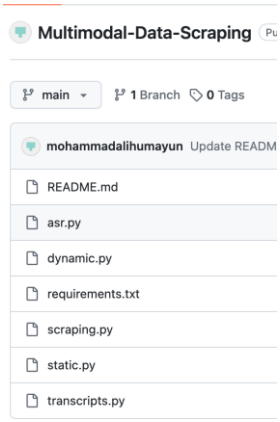



Fig. 8. Github repo structure.

In contrast to existing similar datasets reviewed in earlier sections, our dataset comprises diverse sources, including Articles, Lessons, Lectures, Meetings and Dialogues, Rulings, Speech, Letters, Communications, and Sermons. Additionally, all dataset sources are authentic, sourced from reputable organizations such as the General Authority for the Affairs of the Grand Mosque and the Prophet's Mosque, the Agency of the Affairs of Al-Masjid Al-Nabawi, Sheikh Abdul Aziz bin Baz Charitable Foundation, Sheikh Mohammed bin Saleh Al Othaimeen Charitable Foundation, and the Al Dawa Charitable Foundation. The quality and validity of tools used for speech transcription and data scraping have been thoroughly validated through manual verification of data records. Finally, the automated data scraping method has been confirmed to be reproducible, with clear descriptions of requirements and detailed steps provided for replication of the automated data collection process.

## VI. Conclusions and Limitations

We have developed four distinct algorithms to address different types of text extraction tasks. The first algorithm focuses on extracting text data from static websites. The second algorithm is designed to retrieve text data from dynamic websites. The third algorithm targets the extraction of relevant transcripts from YouTube videos that already have transcripts available. Finally, the fourth algorithm aims to translate speech to text from videos that do not have transcripts. These algorithms enable web scraping [10] across diverse websites containing Islamic knowledge.

Using the proposed web scraping tool, this study has successfully created a comprehensive and reliable Islamic knowledge dataset from authentic Saudi Arabia sources. By aggregating 31,225 records from reputable Islamic scholars, official websites, and authorized YouTube channels, we have established a robust foundation for text processing regarding Islamic knowledge. This dataset addresses the critical need for accurate and contextually relevant Islamic content, aiming for precise and trustworthy responses by automated AI models. The dataset motivates further language processing research for more accurate and useful applications in Islamic knowledge processing.

Admittedly, it was not possible to identify all available Islamic knowledge sources in Saudi Arabia. Therefore, we acknowledge that our dataset might have overlooked other important sources, in which data were not downloadable or unverified. Our Islamic knowledge dataset is collected from sources endorsed by Saudi organizations, thus mainly reflecting the Hanbali school of thought and views. In Islam, there are four main Sunni schools of interpretation endorsed in approximately 30 Muslim countries. The range of topics covered in the proposed dataset is greatly influenced by the content published on the trusted websites that we scraped. Technically, a few limitations caused some sources to be missed. For example, while a time wait was imposed while fetching data from the web elements from dynamic websites, certain elements took a long time to become clickable (i.e., downloadable) and caused a timeout. Moreover, YouTube videos without transcripts were downloaded, converted into audio files, and translated into text. This process was also not foolproof, and some of the videos

failed to download due to network issues or sometimes because an age restriction was set on the target videos, which required a sign-in to download. Finally, our dataset does not include knowledge published in booklets and books, which could hold important insights about Islamic rulings.

## ACKNOWLEDGMENT

This research is supported by the Deanship of Scientific Research of the Islamic University of Madinah, KSA under the research groups (first) project no. 956.

## REFERENCES

[1] A. A. Munshi, W. H. AlSabban, A. T. Farag, O. E. Rakha, A. A. Al Sallab, and M. Alotaibi, 'Towards an automated Islamic fatwa system: Survey, dataset and benchmarks', Int. J. Comput. Sci. Mob. Comput., vol. 10, no. 4, pp. 118–131, 2021.

[2] S. Alnefaie, E. Atwell, and M. A. Alsalka, 'Challenges in the Islamic Question Answering Corpora', Int. J. Islam. Appl. Comput. Sci. Technol., vol. 10, no. 4, pp. 1–10, 2022.

[3] M. Mohammed, S. Amin, and M. M. Aref, 'An english islamic articles dataset (EIAD) for developing an Islambot question answering chatbot', in 2022 5th International Conference on Computing and Informatics (ICCI), IEEE, 2022, pp. 303–309.

[4] M. R. Rizqullah, A. Purwarianti, and A. F. Aji, 'Qasina: Religious domain question answering using sirah nabawiyah', in 2023 10th International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA), IEEE, Oct. 2023, pp. 1–6.

[5] O. Alyemny, H. Al-Khalifa, and A. Mirza, 'A Data-Driven Exploration of a New Islamic Fatwas Dataset for Arabic NLP Tasks', Data, vol. 8, no. 10, p. 155, 2023.

[6] A. M. Farooqi, R. A. S. Malick, M. S. Shaikh, and A. Akhunzada, 'Multi-IsnadSet MIS for Sahih Muslim Hadith with chain of narrators, based on multiple ISNAD', Data Brief, vol. 54, p. 110439, 2024.

[7] M. Mghari, O. Bouras, and A. El Hibaoui, 'Sanadset 650k: Data on hadith narrators', Data Brief, vol. 44, 2022.

[8] A. Namoun, M. A. Humayun, and W. Nawaz, 'Authentic Islamic knowledge dataset'. 2024. doi: 10.17632/zjrc34pc3p.1.

[9] 'Multimodal-Data-Scraping repository'. 2024. https://github.com/anamoun/Multimodal-Data-Scraping.

[10] A. Namoun, A. Alshanqiti, E. Chamudi, and A. Rahmon, 'Web design scraping: Enabling factors, opportunities and research directions', in 2020 12th International Conference on Information Technology and Electrical Engineering (ICITEE), IEEE, Oct. 2020, pp. 104-109.