# Using Web Scraping for Automatic Generation of Structured Arabic Lexicon

**Aya Mohammed Abdul-samad[1], Dr. Salma A. Mahmood[2]**
Department of Computer Information System, College of Computer Science and Information Technology, University of Basrah, Basrah, Iraq.

**Article Info**

***Article history:***

Received December 07, 2023
Revised January 02, 2024
Accepted January 12, 2024

***Keywords:***

Automatic Arabic Lexicon Generation
Web scraping
Web information extraction

**ABSTRACT**

Technological development develops every second increasing text data, especially the Arabic texts on the internet. These Arabic data are massive but it is not useful for use because it is unstructured data and it can’t be used for natural language processing (NLP) and its applications. The increase of Arabic language texts on the Internet has led to an increase in Arabic lexicon web pages but it is not ready for use by NLP applications because it is semi-structured or even unstructured lexicons. The method used in this study is web scraping for scrap data from the internet and converting data from unstructured to structured data. This study aims to build an automatic structured Arabic lexicon ready for NLP and its applications using web scraping. which increases the opportunity to use the Arabic language more widely, which is of great importance in natural language processing applications.

***Corresponding Author:***

Aya Mohammed Abdul-samad
Department of Computer Information System, College of Computer Science and Information Technology, University of Basrah, Basrah, Iraq
Email: itpg.aya.mohammed@uobasrah.edu.iq

## 1. INTRODUCTION

In the expanding field of Natural Language Processing (NLP), languages with rich linguistic diversity and cultural significance are crucial. Arabic, in particular, stands out with its historical importance and complex structure. As technology advances, the importance of Arabic in NLP development becomes evident, presenting challenges and opportunities for researchers. Arabic's intricate grammar, diverse dialects, and contextual script pose unique challenges for NLP applications. Arabic morphology is complex, with roots and patterns playing a key role in word formation. Additionally, the language has a rich system of derivational and inflectional processes, requiring specialized attention from linguists and developers. Developing NLP applications for Arabic is a dynamic field that addresses challenges and takes advantage of unique opportunities. Challenges include capturing the subtleties of Arabic expressions in sentiment analysis and dealing with dialect variability. With the increasing prevalence of digital communication, there is a growing demand for NLP applications that seamlessly integrate with Arabic, enabling users to interact in their native language [1]. In recent decades, there has been an unprecedented surge in the significance and advancement of technology, which has reshaped the manner in which information is accessed, processed, and disseminated. As societies become increasingly interconnected and reliant on digital platforms, the extraction and utilization of data have become of utmost importance. Among the various technological tools that have emerged, web scraping has proven to be highly effective in enabling researchers and practitioners to efficiently collect vast amounts of data from online sources. The proliferation of digital content has brought about a paradigm shift in our approach to information retrieval and content generation. Traditional methods often prove inadequate in dealing with the immense volume and diversity of data available on the web. The development of natural language processing (NLP) technologies has played a pivotal role in this transformative process. NLP, a subset of artificial intelligence, focuses on facilitating machines to comprehend, interpret, and generate human language in a manner that is both contextually accurate and culturally sensitive. Arabic, as one of the world's prominent languages, presents a distinctive set of challenges and opportunities in this technological landscape. The requirement for automated content generation in Arabic is particularly pronounced, considering the dynamic nature of online information and the rich linguistic intricacies of the language. Addressing these challenges necessitates a concerted effort to leverage cutting-edge technologies, such as web scraping, to collect pertinent data, and NLP algorithms to generate content that resonates with Arabic speakers. In this particular context, our research endeavors to explore the potential of utilizing web scraping for automatic content generation in the Arabic language. By doing so, we seek to contribute to the ongoing discourse on the intersection of technology, linguistics, and information retrieval. Through this approach, we aspire not only to advance the field of natural language processing but also to tackle the specific challenges associated with content generation for Arabic speakers in the digital age [2] [3]. In the pursuit of enhancing NLP applications for Arabic, web scraping is a crucial factor. It plays a significant role in enriching linguistic and cultural resources for NLP. Web scraping allows the collection of diverse Arabic-language datasets, including news articles, social media content, and forums. This approach ensures a comprehensive understanding of linguistic variations and idiomatic usage. This study worked to build a large and structured database of Arabic language words and their sections in an automated manner, ready for use by NLP and its applications.

The previous work of this study Batarfi et al. (2019) This investigation entails the construction of an automated Arabic semantic lexicon through the selection of an Arabic lexicon and the augmentation of morphological and semantic data. The objective of this inquiry is to develop a straightforward approach to extracting lexical entries and establishing associations among words [3]. Jarrar et al. (2019) The paper commences with an examination of the criteria utilized to depict linguistic data. Furthermore, it delves into the realm of digital lexicographic resources and the recommendations put forth by The World Wide Web Consortium (W3C) with regard to linguistic data sources. It acknowledges the existence of initiatives such as Linguistic Linked Open Data Cloud (LLOD) . Ultimately, it emphasizes the dearth of Arabic lexicons and past endeavors to illustrate Arabic morphological lexicons [4]. Subhan et. al (2019) This investigation centers on the execution of a semantic examination of the data derived from the social networking platform Twitter, with the purpose of gathering organized data on traffic congestion. The investigation encompasses the phases of data compilation and analysis, uncovering linguistic elements and their denotations, as well as segregating data regarding the occurrence of congestion in terms of both location and time, based on the publication of tweets on Twitter [5]. Alexandrescu et al. (2019) This paper describes the creation of a specialized search engine that retrieves specific information. The process includes retrieval, extraction, presentation, and delivery. Data is collected from web pages and relevant information is extracted and organized for user-friendly presentation. The search engine allows efficient access to information. The paper introduces innovative techniques for each stage of the process, including representing templates, optimizing product indexing, and deploying data in a cloud-based database. Ethical and security considerations are also discussed. A practical example is provided, and suggestions for further research are given to improve cost efficiency and scalability. In summary, the paper details the steps and introduces innovative methods for building search engines, emphasizing efficient resource usage and ethical considerations [6]. Khder et al. (2021) The paper offer a detailed analysis of web scraping or web crawling, which involves extracting data from websites using software. Search engines like Google were among the first to develop scrapers that scan and index web pages. Web scraping mimics human browsing through web browsers or HTTP. It can convert unstructured web data to structured data [7]. Egger et al. (2022) this paper on the impact of digitalization on the tourism industry and underscores the indispensability of collecting and analyzing data for a more profound comprehension of customers, competitors, and stakeholders. Additionally, the paper underscores the pivotal role of web scraping in gathering and retrieving data from the internet and furnishes information on the available tools and packages for web scraping [8]. Brenning et al. (2023) The present paper offers geographical research studies that employ the technique of web scraping, demonstrating its applicability across diverse research domains. The paper accentuates the benefits of web scraping, encompassing instantaneous access to geographically located data and cost-effectiveness in contrast to conventional data acquisition approaches. Additionally, it delves into the obstacles posed by web scraping, which encompass ethical and legal dilemmas pertaining to intellectual property rights, informed consent, and confidentiality, as well as technical challenges embracing data inconsistencies and partiality [9]. Sirisuriya et al. (2023) Machine learning algorithms require large amounts of data to make automatic predictions, and data scraping solves this problem of collecting large amounts of data by collecting data from web pages and extracting data from them that are usually unstructured or semi-structured and transforming them into structured and organized data, the collected data is important for machine learning algorithms [10]. Shreekumar et al. (2023) In this paper, the importance of data scraping is defined as an indispensable tool for electronic commercial companies. Data scraping plays an important role in the process of collecting information, which enables important decisions that depend on data scraped from the web [11]. ‘

The paper is arranged as follows. Section 2, introduces the methodology of how to build an Arabic lexicon. Section 3, results and the discussion of the proposed method. Finally, we address the conclusions in Section 4.

## 2. METHODOLOGY

The main problem of this study is how to build an automatically structured Arabic lexicon using web scraping. The following figure is the proposed framework:

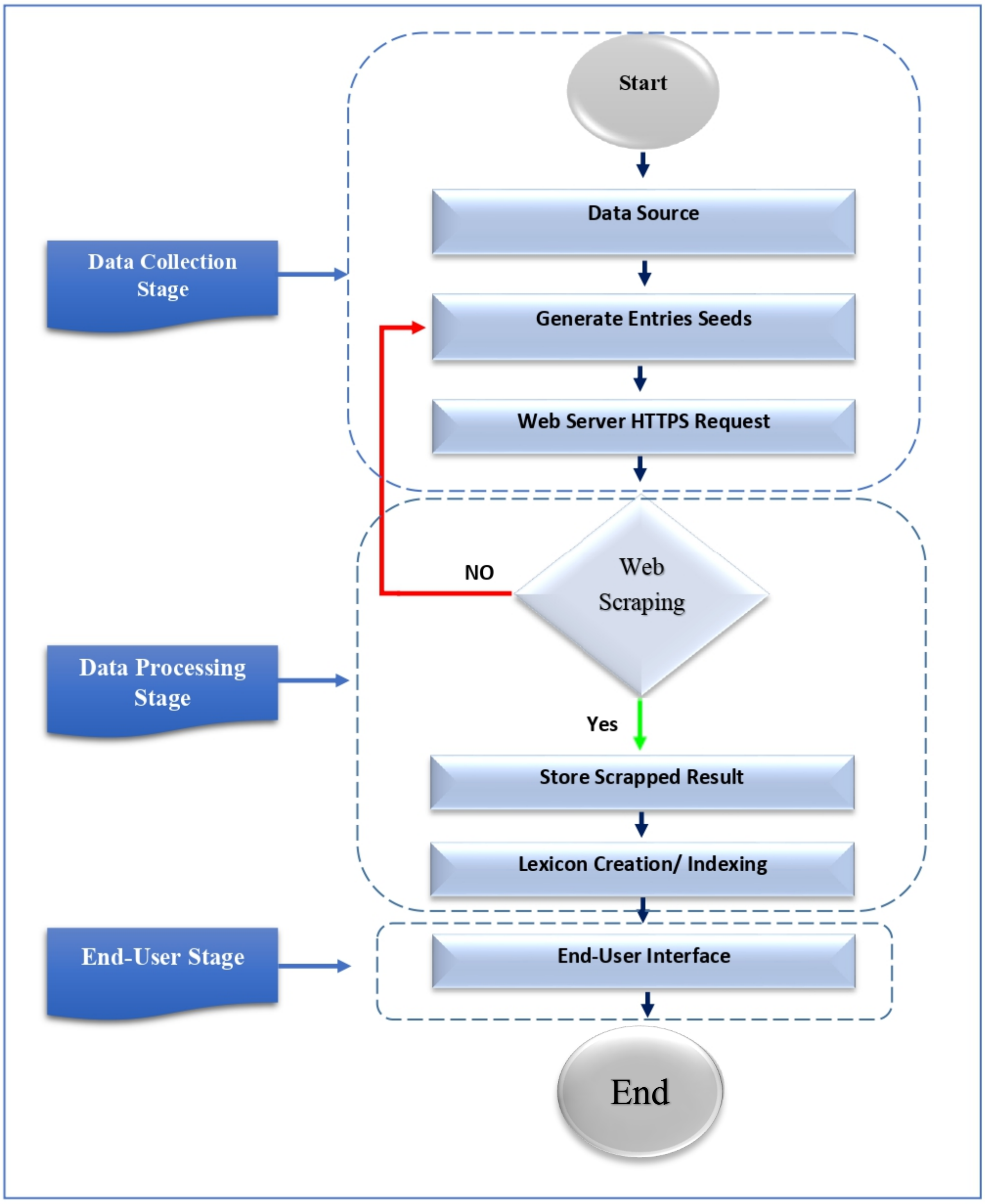


**Figure 1: Proposed Framework**

The diagram Figure 1 represents an explanation of the basic steps of the methodology for building an automated, structured Arabic lexicon**.**

### 2.1 Data Collection Stage

The data collection stage begins with choosing the data source from which the data will be extracted, i.e. the web page chosen in order to extract the data from, which must meet a set of standards after which we can guarantee the accuracy and credibility of the lexicon that will be collected. The lexicon must be comprehensive and highly accurate in selecting its sources. It must be diverse in fields, open source in order to avoid the ethical laws of the data scraping process, and contain morphological, grammatical, and semantic information, in addition to the continuous updating of the lexicon to keep pace with developments in the language. The almaany lexicon was chosen because it meets almost all the requirements that achieve the goal [13].

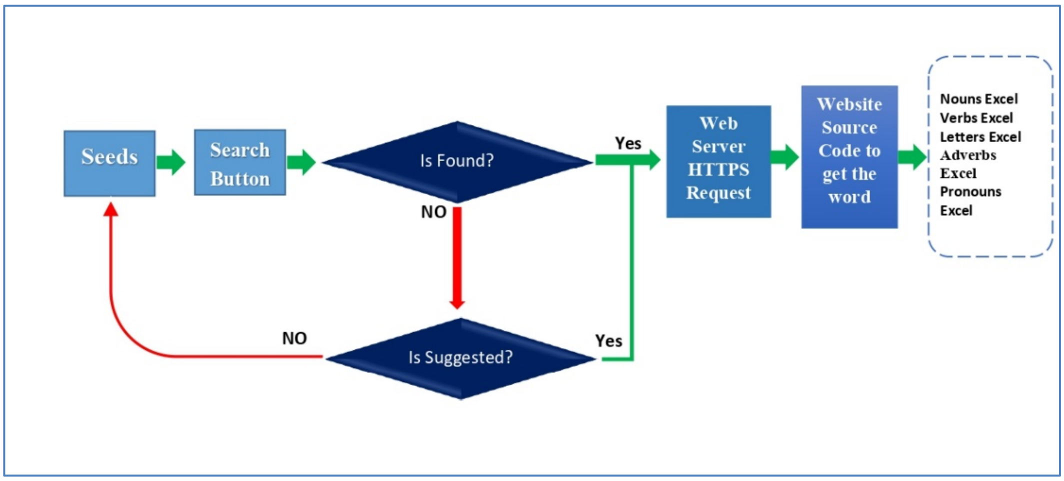
The next stage in the data collection process is the process of preparing seeds (words sent to the lexicon web page to be retrieved and scraped from the web page). The website that was chosen to scrape data from is dynamic, that is, it presents private content for each user. This content cannot be scraped unless we have seeds sent to the web page and that word is retrieved so that it can be scrapped. These seeds were obtained from a static web page, meaning that it displays the same content to all users. This type of web page is easier to scrap from dynamic web pages. For this reason, approximately two hundred thousand words were collected from seeds [14].

## 2.2 Data Processing Stage

Web scraping is a method employed to extract data from websites. This procedure encompasses retrieving the webpage and subsequently extracting and parsing the desired information. Although web scraping can be accomplished manually, it is frequently executed through the utilization of automated tools or bots.

This stage is one of the most important stages on which the construction of the dictionary depends. It is the process of scraping data from the web and structuring it to be ready for use by natural processing language applications. This process is done by sending seeds to the lexicon web page, and the best way to access the pages for words is through the search box on the website. The search box serves as a window to enter the website's database. The seeds are sent one by one to the search box to retrieve the page on which the searched word is located.

In Figure 2, after placing the seed in the search field, the word will go through several processing operations to scrape it.



.

**Figure 2: Web Scraping Process**

When sending the single seed to the web page, work is done to match the word with the content present in the lexicon’s web page in terms of whether this seed exists or not. If it is found, it goes to the next step. If it does not exist, the system will work by suggesting words that are close to the original word. To be taken instead if it does not exist, but if it is not found and there is no suggested word to replace it, the system automatically moves to the next seed to perform the same steps on it. After verifying that the word is in the lexicon, go to the page where the word is found. It is known that speech in the Arabic language consists of several parts, including nouns, verbs, letters, pronouns, and adverbs. On the page where the word is found, there are several parts of speech for the word. After that, the code of the lexicon’s web page is accessed and the word is then scraped from the web, not only is the word scraped, but the representations of this word (its meaning, root, context, example, synonym, opposite... etc.) are scraped along with the word and exported to Excel for the word according to its classification from the parts of the Arabic language. For example, if the word is a verb, all related representations of the word are scraped. When the word comes as a subject or object, its type is scraped, whether it is an intransitive verb or a transitive verb, in addition to the meaning, context, example, root, synonym, and opposite of the word. They are scraped and arranged in the verb Excel, meaning the structuring and classification are done during Scraping. The entries for the Arabic language on which all the words that were scraped were classified were (nouns, verbs, letters, adverbs, and pronouns), that is, almost all parts of speech in the Arabic language were included, and each table contains its structure, meaning that the table of nouns differs from the table of verbs. Therefore, they differ from the table of letters, pronouns, and adverbs. This method is the best method for web scraping.

### 2.3 Lexicon Creation

The Arabic lexicon is of great importance in the field of information technology and computer science because of its importance in natural language applications such as sentiment analysis, information retrieval and other applications that are of great importance to many institutions and organizations. This lexicon is also considered a database for machine learning.

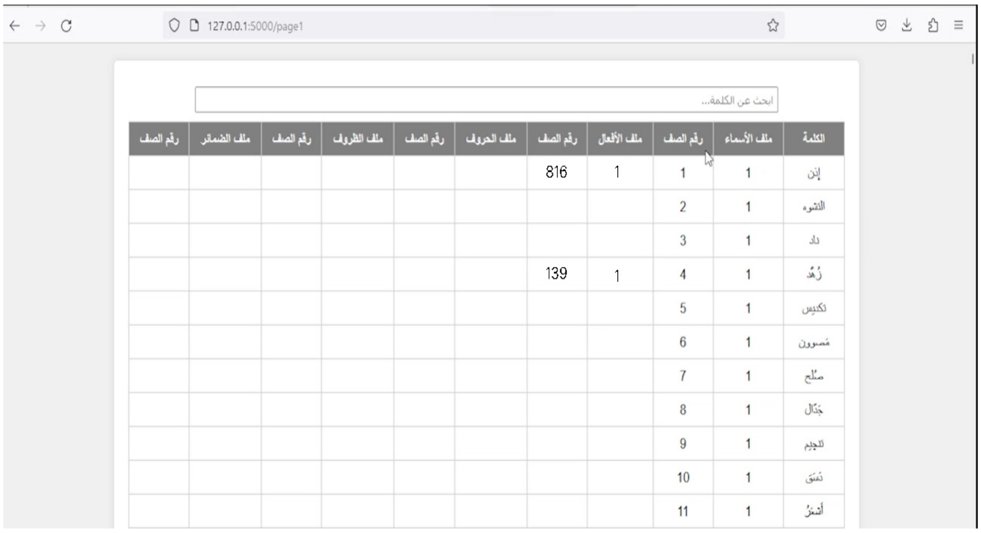
### 2.3.1 The Indexing

The process of gathering input files in order to construct a lexicon by means of indexing, with the intention of enhancing the ease of word retrieval, involves scrutinizing each individual word to ascertain its classification and subsequently assigning the corresponding class number to said word.

.

### 2.3.2 Joint Indexing Mechanism

It is known that the Arabic language is a morphological and semantic language, and a single word can be a noun and a verb at the same time, or it can be a noun, an adverb, and a pronoun at the same time. The same word is scratched out, but it is classified according to the five entries that we mentioned. A joint indexing system has been implemented for words that share more than one table, indicating the row number and name of the table in which they are located, as shown in the following figure:



127.0.0.1:5000/page1

ابحث عن الكلمة...

| رقم الصف | ملف الضمائر | رقم الصف | ملف الظروف | رقم الصف | ملف الحروف | رقم الصف | ملف الأفعال | رقم الصف | ملف الأسماء | الكلمة |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | 816 | 1 | 1 | 1 | إذن |
| | | | | | | | | 2 | 1 | التشوه |
| | | | | | | | | 3 | 1 | داد |
| | | | | | | 139 | 1 | 4 | 1 | زُهْد |
| | | | | | | | | 5 | 1 | تكبيس |
| | | | | | | | | 6 | 1 | مَصوون |
| | | | | | | | | 7 | 1 | صُلح |
| | | | | | | | | 8 | 1 | جَدّال |
| | | | | | | | | 9 | 1 | تلجيم |
| | | | | | | | | 10 | 1 | دَنَق |
| | | | | | | | | 11 | 1 | أشعَرُ |

**Figure 3: Joint Indexing**

The number (1) was placed in the entry name field so that we know that this word belongs to this entry, as well as the number of the field in which the word is located, about the table in which it is found.

### 2.4 End- User

Creating a lexicon interface has several advantages, such as providing an easy way for users to interact with the lexicon. Users can search and access the lexicon content easily without technical skills or a user manual. Interfaces connect users and programming data, making interaction with the lexicon easier. Flask framework is used for building user interfaces with Python, and for efficient searching and indexing in the lexicon a Trie Data Structure is used [15].

### 2.5 Legal Web Scraping

When web scraping, it is important to consider legal and ethical issues and avoid violating website rules. Illegal scraping can result in legal consequences, including punishment. Copyright infringement should be avoided. Server blocking can also lead to legal problems, even if scraping is allowed by the website, as it can be seen as a form of hacking [12].

## 3. RESULTS AND DISCUSSION

The web scraping process yielded a comprehensive Arabic lexicon with the following key statistics:

Table 1: Entrance Statistics

| Entrances | Number of Words Collected from Web Scraping |
|---|---|
| Nouns | 10000 |
| Verbs | 10000 |
| Letters (Prepositions, Conjunctions) | 70 |
| Adverb | 500 |
| pronouns | 20 |

As indicated in the table (1), the Arabic language consists of several basic parts, as shown in the table above. each of these parts represents a basic entrance in arabic language. These entrances each have their own function according to the part of speech in which they fall, according to the sentences. Each entry has been separated individually to be organized and structured. Each entrance has its own unique characteristics that differ from other entrances.

The goal of this study is to build a structured, comprehensive, and integrated lexicon that covers many of the morphological, semantic, and other aspects that were covered. This dictionary was built to be ready for NLP and its applications such as (sentiment analysis, information retrieval, and translators).

## 4. CONCLUSION

The structured and automated Arabic lexicon has been completed using web scraping. Various challenges were faced and although difficult, the lexicon is satisfactory and valuable. It addresses all the complexities and gaps in understanding the Arabic language. Strategies have been developed to overcome the linguistic complexity of Arabic and appreciate the cultural and historical roots of the language. The lexicon is divided into different categories such as nouns, verbs, letters, pronouns, and adverbs. It is now ready for automated processing in NLP applications. The essence of this project lies in skillfully extracting data from the web and using effective retrieval methods, which opens doors to many applications.

## REFERENCES

[1] H. Ishkewy, H. Harb, and H. Farahat, “Azhary: An Arabic Lexical Ontology,” 2014.

[2] د. ع. صابر, د. ع. مرهون, د. س. ع. محمود, and “تصميم نظام استخلاص المعلومات من بعض فقرات النصوص العربية,” Journal of Education for Pure Science, vol. 1, no. 2, 2010.

[3] O. Batarfi, M. Yehia, and A. Ezz, “Building an Arabic semantic lexicon for hajj,” Int. J. Comput. Appl., vol. 181, no. 39, pp. 9–15, 2019.

[4] M. Jarrar and H. Amayreh, “An Arabic-multilingual database with a lexicographic search engine,” in Natural Language Processing and Information Systems, Cham: Springer International Publishing, 2019, pp. 234–246.

[5] S. Subhan, E. Sediyono, and F. Farikhin, “The semantic analysis of Twitter Data with Generative Lexicon for the information of traffic congestion,” Journal of Advances in Information Systems and Technology, vol. 1, no. 1, pp. 45–54, 2019.

[6] A. Alexandrescu, “Optimization and security in information retrieval, extraction, processing, and presentation on a cloud platform,” Information (Basel), vol. 10, no. 6, p. 200, 2019.

[7] M. Khder, “Web scraping or Web Crawling: State of art, techniques, approaches and application,” Int. J. Adv. Soft Comput. Appl., vol. 13, no. 3, pp. 145–168, 2021.

[8] R. Egger, M. Kroner, and A. Stöckl, “Web scraping: Collecting and retrieving data from the web,” in Applied Data Science in Tourism, Cham: Springer International Publishing, 2022, pp. 67–82.

[9] A. Brenning and S. Henn, “Web scraping: a promising tool for geographic data acquisition,” arXiv [cs.IR], 2023.

[10] S. D. S. Sirisuriya, “Importance of web scraping as a data source for machine learning algorithms - review,” in 2023 IEEE 17th International Conference on Industrial and Information Systems (ICIIS), 2023.

[11] S, S, S.M, and M.D, "IMPORTANCE OF WEB SCRAPING IN E-COMMERCE BUSINESS."2023

[12] A. Luscombe, K. Dick, and K. Walby, “Algorithmic thinking in the public interest: navigating technical, legal, and ethical hurdles to web scraping in the social sciences,” *Qual. Quant.*, vol. 56, no. 3, pp. 1023–1044, 2022.

[13] Almaany team, "قاموس عربي عربي وقاموس عربي انجليزيقاموس ومعجم المعاني متعدد اللغات والمجالات ثنائي" almaany.com. https://www.almaany.com/(accessed may. 22, 2022)

[14] H. Petersen, K. Blum, T. Tamme, and M. Tartu, “From static and dynamic websites to static site generators,” *Core.ac.uk*. [Online]. Available: https://core.ac.uk/download/pdf/83597655.pdf.

[15] M. R. Mufid, A. Basofi, M. U. H. Al Rasyid, I. F. Rochimansyah, and A. Rokhim, “Design an MVC model using python for flask framework development,” in 2019 International Electronics Symposium (IES), 2019.