data

MDPI

Article

A Data-Driven Exploration of a New Islamic Fatwas Dataset for Arabic NLP Tasks

Ohoud Alyemny [1,*], Hend Al-Khalifa [2] and Abdulrahman Mirza [1]

[1] Department of Information Systems, College of Computer and Information Sciences, King Saud University, Riyadh 12372, Saudi Arabia; amirza@ksu.edu.sa

[2] Department of Information Technology, College of Computer and Information Sciences, King Saud University, Riyadh 12372, Saudi Arabia; hendk@ksu.edu.sa

* Correspondence: oalyemni@ksu.edu.sa

Abstract: Islamic content is a broad and diverse domain that encompasses various sources, topics, and perspectives. However, there is a lack of comprehensive and reliable datasets that can facilitate conducting studies on Islamic content. In this paper, we present *fatwaset*, the first public Arabic dataset of Islamic fatwas. It contains Islamic fatwas that we collected from various trusted and authenticated sources in the Islamic fatwa domain, such as agencies, religious scholars, and websites. *Fatwaset* is a rich resource as it does not only contain fatwas but also includes a considerable set of their surrounding metadata. It can be used for many natural language processing (NLP) tasks, such as language modeling, question answering, author attribution, topic identification, text classification, and text summarization. It can also support other domains that are related to Islamic culture, such as philosophy and language art. We describe the methodology and criteria we used to select the content, as well as the challenges and limitations we faced. Additionally, we perform an Exploratory Data Analysis (EDA), which investigates the dataset from different perspectives. The results of the EDA reveal important information that greatly benefits researchers in this area.

Keywords: fatwas; exploratory data analysis; Islamic content; natural language processing

check for updates

Citation: Alyemny, O.; Al-Khalifa, H.; Mirza, A. A Data-Driven Exploration of a New Islamic Fatwas Dataset for Arabic NLP Tasks. *Data* **2023**, *8*, 155. https://doi.org/10.3390/data8100155

Academic Editor: Maarten Marx

Received: 24 August 2023
Revised: 28 September 2023
Accepted: 17 October 2023
Published: 19 October 2023

1. Introduction

Islam is characterized by an extensive body of unchanging laws (called *Shari'ah*). There are two primary sources of Islamic rules: the *Quran* (the holy book of Islam) and *sunnah* (authenticated prophetic teachings) [1,2]. Muslims' daily lives and social dealings are highly influenced by Islamic rules. Accordingly, this raises a lot of questions about specific issues that require answers or *Fatwas*. *Fatwa* refers to the formal rulings issued by a qualified scholar of Islamic law (*mufti*) [3]. Specifically, it is a mufti's answer to a question regarding Islamic laws [4].

Given the central significance of *fatwa* to the Islamic community and non-Muslim knowledge-seekers, there is a need to benefit from the advances in computational techniques to support this area. More specifically, it is necessary to utilize the improvements in the field of Natural Language Processing (NLP) to facilitate Islamic content research and studies. While a dataset is usually the main roadblock in NLP, datasets in the Islamic content domain have more unique challenges. The nature of the sources of content poses difficulties when constructing and annotating datasets. Some of the resources are only available in books and as recorded audio. Others are scattered online among websites that take different formats. The annotators must refer to authorized, authenticated, and trusted resources to label the data. The preprocessing phase is very time consuming due to the different structures of the content and the challenges of the Arabic language. Along with this, dataset availability is an open issue in the Islamic content field, as there are a limited number of available datasets. Reviewing the literature showed that most Islamic content datasets are dedicated to the Quran. On the other hand, other types of Islamic content, such as *Sunnah* and *fatwas,* have not received considerable interest [5]. In that regard, there is no public dataset dedicated to *fatwas* in Arabic.

Therefore, this paper introduces *Fatwaset*, which is a dataset for Islamic fatwas that covers various topics and includes fatwas from several authenticated and trusted agencies, religious scholars, and websites. The dataset does not only contain fatwas but also incorporates important metadata for each fatwa. *Fatwaset* is a rich and fruitful source of knowledge for researchers that fosters research on Arabic NLP in general and Islamic content in particular. It supports the scholarly community for a wide range of NLP problems. The following illustrate some of the use cases:

- *Fatwaset* is a large and diverse Arabic dataset that covers Islamic text from several Arab countries. This makes it suitable for training Arabic language models and domain-specific language models. For instance, pre-training a language model using *Fatwaset* can help in building effective systems that detect anti-Islamic or hateful content from social media platforms.
- It can be augmented in a chatbot system to answer questions about Islamic content. For instance, *Fatwaset* can be used to build a question answering system that provides answers to queries about Islamic topics, such as " What are the five pillars of Islam?".
- It is an excellent option for author attribution tasks. It contains a great number of texts from a considerable set of authors, which makes it possible to train author attribution models. For instance, *Fatwaset* allows for training a model that learns the features and patterns of religious scholars' answers in terms of vocabulary, style, and structure. Then, this model can be tested to identify the religious scholar of a new given text. Also, it can be used to evaluate and compare the effect of several features in an author attribution task.
- Because it provides a considerable set of metadata for each fatwa text, it can be used in topic identification, clustering, and text classification tasks. For instance, each fatwa in *Fatwaset* has a title that allows for building models that are able to cluster Islamic texts into groups based on their similarity in terms of topics and vocabulary.
- It contains long texts that can be used in text summarization tasks; for instance, *Fatwaset* contains a large and diverse collection of answers that support training models that generate abstractive coherent summaries from religious scholars' answers.
- It can be used and extended to support other domains, such as philosophy, history, language art, and social science, due to its strong connection with Islamic spiritual culture.
- To our knowledge, *Fatwaset* is the first available Arabic dataset for Islamic fatwas.

Additionally, this paper conducts an Exploratory Data Analysis (EDA) to study *fatwaset* from different angles. The results of the EDA open up avenues for further areas of study and investigation. Thus, our objectives in this paper are the following:

1. Construct *Fatwaset*, the first public dataset of Islamic fatwas in Arabic, to enable researchers in computational linguistics to conduct studies on Arabic and Islamic-related NLP problems.
2. Understand the content of Islamic fatwas by performing an Exploratory Data Analysis on *Fatwaset*.

2. Background and Related Works

2.1. Islamic Fatwa-Related Studies

Islamic content, in its many and varied forms, plays a fundamental role in Muslims' everyday lives. Much of the interest in Islamic content comes from its enormous importance to the Islamic community and the realization of its interdisciplinary nature. It attracts scholars from various fields, such as philosophy, history, art, literature, and politics. While a dataset is a critical enabler and an essential component in conducting studies, the number of available datasets in the Islamic content domain is very limited.

In the context of datasets dedicated to the Quran, Malhas and Elsayed constructed AyaTEC [6], which is the gold standard Arabic corpus. It has been used as a test collection for a shared task about verse-based question answering over the Quran. It includes

207 questions (with their corresponding 1762 answers, as some questions have more than one answer). The corpus covers 11 different topic categories. It is worth mentioning that the questions cover only 25% of the Quran verses [5].

Moreover, the Quranic Reading Comprehension Dataset (QRCD) is an extension of the AyaTEC dataset [7]. It includes 1093 pairs of Quranic passages (set of verses) and a corresponding question on the passage. It also includes 1337 answers (some questions have more than one corresponding answer). Here, the dataset covers about 53% of the Quran [5].

There are also some datasets that are intended to serve types of Islamic content other than the Quran. For instance, EIAD [8] is a dataset of English Islamic articles that contains about 10,000 English Islamic articles pulled together from three authenticated websites. The dataset is categorized into 15 different categories. Each category includes several distinct topics. The authors built the dataset to develop a chatbot for answering questions about Islam. The intended users are new Muslims and non-Muslims who want to convert to Islam.

A considerable dataset is the one introduced by AlZahrani and Al-Yahya, which consists of Islamic legal texts with their corresponding authors for the task of authorship attribution. The dataset is in Arabic and was extracted from books. It offers 200 texts from 40 authors [9].

Apparently, there is only one paper that focuses on constructing a dataset for fatwas, which is introduced in [1]. The authors worked on a dataset for Islamic fatwas in Arabic. The dataset contains about 850,000 fatwas collected from 11 trusted websites. The dataset only includes data about questions, answers, fatwa topics, and publication dates. It does not contain other metadata. Metadata is defined as "data about data" [10]. It gives information that assists in describing, understanding, and managing the content of a resource [11]. However, the dataset is not publicly available.

After a closer examination of the available datasets in the Islamic content domain, we want to shed light on the following points:

1. In the literature, the focus has been mainly placed on Quran datasets. In contrast, other types of Islamic content datasets such as *Sunnah* and *fatwa* have not received considerable attention [5].
2. Currently, most of the available datasets are designed specifically to target question answering tasks. This limitation in design restricts expansion of the pool of Islamic content-related studies. Dataset design should incorporate criteria that facilitate undertaking other tasks.
3. It is evident that Arabic is the dominant language across Islamic content datasets. This might be because the original resources are available in Arabic. However, there is a need to address other languages.
4. Regarding datasets concerning fatwa, there is only one dataset presented in [1]. However, the dataset is not publicly available. Conversely, the proposed dataset in this paper, *fatwaset*, is public to the academic community.
5. The fatwa dataset introduced in [1] only includes data about fatwa questions, fatwa answers, fatwa topics, and publication date. It does not contain other metadata about the fatwas. The inclusion of other metadata enriches the dataset and increases its usability. Therefore, *fatwaset* includes all the metadata related to a certain fatwa in the given resource (the collected metadata is presented in Section 3.1). The aim is to make the dataset effective and applicable to studying a range of Natural Language Processing and text mining problems.

2.2. Exploratory Data Analysis (EDA)

EDA is mainly a technique that summarizes the dataset by extracting its main characteristics and visualizing it using appropriate representations [12]. It is an important step after preprocessing the collected data to understand, visualize, and manipulate the data without making any assumptions. EDA also assists analysts in gaining deep insights into the dataset to identify potential relationships between variables and find abnormalities and outliers [13,14].

EDA techniques can be classified into graphical methods (such as histograms, heatmaps, and side-by-side boxplots) and non-graphical methods (such as covariance and correlation measures) [14]. One of the most common graphical techniques in EDA is the word cloud. A word cloud is a visual representation of word frequency. If a word appears frequently in textual data, it will be represented in a larger font size in the word cloud. The word cloud technique has been used in several domains, especially in content management systems and crowdsourcing platforms [15].

In a recent study, Balz utilized EDA techniques to analyze all papers published in MDPI's *Remote Sensing* journal between 2009 and 2021. The analysis covers the full text of 19,289 articles. The author uses word clouds to visualize the research topics and trends. The text analysis shows that each sub-field of remote sensing has its own special style and writing patterns in publications. Moreover, it showed distinctive writing styles for papers from different countries and even cities [16].

Additionally, the authors in [17] presented the Saudi Novels Corpus and used the capabilities of EDA to analyze the collected data. The study shows the top ten unigrams, bigrams, and trigrams, which have been compared with another corpus. The study also reported the top part-of-speech tags.

In this paper, we conduct a detailed EDA on the collected dataset, *fatwaset*. We perform analysis on several of its aspects. The results of the EDA carry extensive information that can be of great value to interested researchers.

3. Materials and Methods

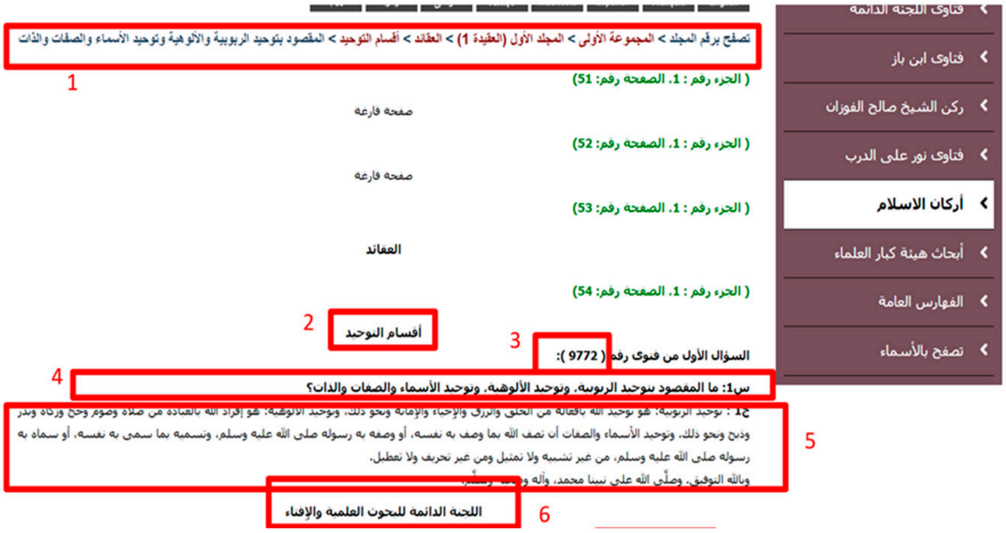
3.1. *Fatwaset*

We trawled 13 popular websites on Islamic fatwas that are available in Arabic. The selection of the websites focuses on covering several geographical areas in the Arab world. The aim is to try to cover and capture more topics and different linguistic terms for the purpose of enriching our corpus. This is because each country has its own culture, which implies the usage of certain words and the discussion of specific topics that rarely appear in other cultures and countries. The dataset includes websites that represent official agencies supported by governments (such as Dar Al Ifta in Saudi Arabia), official websites that belong to famous religious scholars (such as Alshaikh Abdual Aziz Ibn Baz's website), and other trusted popular websites in the field (such as Islamweb). The dataset includes fatwas from the websites listed in Table 1.

Table 1. List of the used websites for collecting *fatwaset* [accessed 26 May 2023].

Website	Link	Country
Dar Al Ifta in Saudi Arabia	https://www.alifta.gov.sa/	Saudi Arabia
Dar Al Ifta in Egypt	https://www.dar-alifta.org	Egypt
Dar Al Ifta in Jordan	https://aliftaa.jo	Jordan
Al Shaikh Abdual Aziz Ibn Baz	https://binbaz.org.sa	Saudi Arabia
Al Shaikh Mohammad Ibn Othaimin	https://binothaimeen.net/site	Saudi Arabia
Al Shaikh Abdual Aziz Al Ashaikh	https://www.mufti.af.org.sa	Saudi Arabia
Al Shaikh Saleh Al Fwzan	https://www.alfawzan.af.org.sa	Saudi Arabia
Al Shaikh Saleh Bin Humaid	https://www.ibnhomaid.af.org.sa/	Saudi Arabia
Al Shaikh Abdullah Al Manee	https://al-manee.com	Saudi Arabia
IslamWeb	https://www.islamweb.com	Qatar
FatwaPedia	https://fatawapedia.com	Saudi Arabia
IslamQA	https://islamqa.info	Syria
IslamOnline	https://islamonline.net	Qatar

During the collection, we also collected the metadata of the fatwas. In other words, the dataset does not only contain fatwas (questions and answers), it also includes all the other given data on the page related to the fatwa (e.g., date of publication, category, title, subtitle, name of scholar, etc.). These fields can be considered metadata. The aim is to make the dataset effective and appropriate for studying a variety of issues. For instance, it would be possible to study the changes in fatwas asked by Muslims over time (in terms of topics, details, and quantity) or cluster fatwas according to a specific variable. As an example, Figure 1 illustrates a fatwa and the metadata collected from the Dar Al Ifta website in Saudi Arabia.



تصفح برقم المجلد > المجموعة الأولى > المجلد الأول (العقيدة 1) > العقائد > أقسام التوحيد > المقصود بتوحيد الربوبية والألوهية وتوحيد الأسماء والصفات والذات

(الجزء رقم : 1، الصفحة رقم: 51)

صفحة فارغة

(الجزء رقم : 1، الصفحة رقم: 52)

صفحة فارغة

(الجزء رقم : 1، الصفحة رقم: 53)

العقائد

(الجزء رقم : 1، الصفحة رقم: 54)

أقسام التوحيد

السؤال الأول من فتوى رقم (9772):

س1: ما المقصود بتوحيد الربوبية، وتوحيد الألوهية، وتوحيد الأسماء والصفات والذات؟

ج1 : توحيد الربوبية: هو توحيد الله بأفعاله من الخلق والرزق والإحياء والإماتة ونحو ذلك، وتوحيد الألوهية: هو إفراد الله بالعبادة من صلاة وصوم وحج وزكاة ونذر وذبح ونحو ذلك. وتوحيد الأسماء والصفات أن تصف الله بما وصف به نفسه، أو وصفه به رسوله صلى الله عليه وسلم، وتسميه بما سمى به نفسه، أو سماه به رسوله صلى الله عليه وسلم، من غير تشبيه ولا تمثيل ومن غير تحريف ولا تعطيل.

وبالله التوفيق، وصلّى الله على نبينا محمد، وآله وصحبه وسلّم.

اللجنة الدائمة للبحوث العلمية والإفتاء

Figure 1. Data collected from "Dar Al Ifta in Saudi Arabia" website. 1. Main title (the meaning of oneness of lordship, worship, names, attributes and self), 2. subtitle (types of oneness), 3. fatwa number, 4. fatwa question (what is the meaning of oneness of lordship, worship, names, attributes and self?), 5. fatwa answer (oneness of lordship: is oneness of Allah (God) through His actions of creation, provision, life, death, and so on. Oneness of worship is the worship of Allah alone in prayer, fasting, Hajj, zakat, vows, sacrifice, and so on. Oneness of names, attributes: is describing Allah the way He has described Himself, and the way His Messenger (peace be upon him) described Him; and naming Him with the Names that He has named Himself with, and His Messenger (peace be upon him) named Him with, without comparison, likening Allah's Attributes to those of His Creation, allegorical interpretation nor denial of Allah's Attributes), 6. mufti name.

During data collection, we faced several challenges illustrated in the following:

1. The different formats of each website were a great challenge. For instance, each website has its own way of categorizing fatwas. Some websites classify based on *Figh* (jurisprudence) and subject categories, while others use main topics and subtopics. There are also some websites that do not organize fatwas into categories; fatwas are just posted in a list without any order;
2. The number of given metadata related to fatwas is not the same for each website because each one provides a different set of metadata;
3. Some websites replace the text of the answer with an audio clip;
4. The problem of different used hierarchies for pages from the same website.

Accordingly, these problems required more time and effort to understand each website format and modify the programming code to adapt to the changes and solve these problems.

The number of collected fatwas for each website is represented in Table 2. Each record contains a fatwa question, fatwa answer, and the corresponding metadata (according to the selected website). It is worth mentioning that the given number of rows is after the cleaning process, which includes removing duplication and rows that have empty fields (more than 55,000 rows have been removed during the cleaning process).

Table 2. Number of collected fatwas and metadata for each website.

Website	Number of Records	Metadata
Dar Al Ifta in Saudi Arabia	20,000	Main title, Subtitle, Fatwa number, Mufti name
Dar Al Ifta in Egypt	3769	Main title, Subtitle, Fatwa number, Publication date, Mufti name
Dar Al Ifta in Jordan	3146	Main title, Subtitle, Fatwa number, Publication date, Mufti name
Al Shaikh Abdual Aziz Ibn Baz	27,111	Fighi main title, Subject main title, Subtitle
Al Shaikh Mohammad Ibn Othaimin	9125	Main title, Subtitle, Fatwa number
Al Shaikh Abdual Aziz Al Ashaikh	135	Title
Al Shaikh Saleh Al Fwzan	1723	Title, Answer source
Al Shaikh Saleh Bin Humaid	48	Title
Al Shaikh Abdullah Al Manee	233	Title
IslamWeb	11,662	Main title, Subtitle, Fatwa number, Publication date
FatwaPedia	47,098	Title, Mufti name
IslamQA	195	Title, Answer summary
IslamOnline	5937	Main title, Subtitle
Total	**130,182**	

3.2. Proposed Pipeline of Exploratory Data Analysis (EDA)

The first phase of the pipeline after reading the dataset is preprocessing, which involves a set of steps. First is tokenization, where the text is split into a list of tokens (based on spaces). Next, punctuation marks have been removed. Because the text is in Arabic and some words have diacritics, there was a diacritic removal step. Usually, in Natural Language Processing (NLP), it is necessary to remove the stop words, which are words that occur commonly in the language (such as prepositions, pronouns, etc.). This is because these words do not carry much information. Also, the high frequency of these words increases the size of the dataset and makes computation slower. The stop words we removed are from the list proposed in [18]. Additionally, we removed some unnecessary words that are explicitly added in the Islamic fatwa domain, such as: السؤال(Alsuwal) the question, الجواب(Aljawab) the answer, الفتوى(Alfatwaa) the fatwa, الشيخ(Alshaykh) the religious scholar.

We used several Python libraries to preprocess and visualize the data. For data manipulation, we used Pandas and Numpy. For data visualization, we used Matplotlib and Wordcloud. The collections library has been used to retrieve the counter for each word. We have also used the Natural Language Toolkit (nltk) to tokenize the data. Because the text is in Arabic, we had to use special libraries such as pyarabic to remove diacritic marks and arabic_reshaper to reconstruct the words to be used in word clouds.

Then, we started to investigate the dataset to find valuable information and representations. We focused on three aspects: fatwas' topics, fatwas (questions and answers), and the way religious scholars answer fatwas' questions (length of given answers).

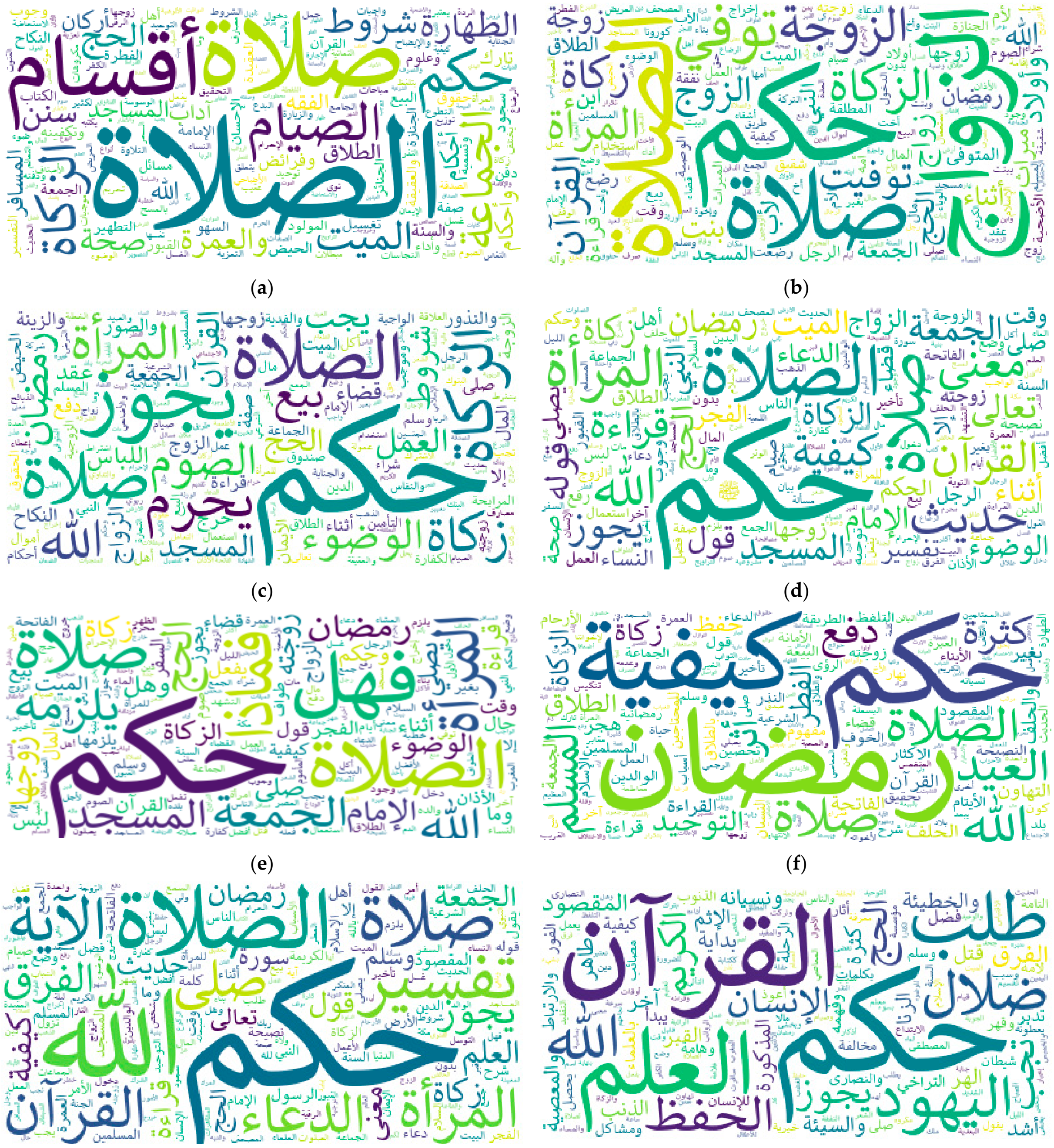
4. Results and Discussion

4.1. Fatwas' Topics

As our dataset offers the titles of fatwas, we worked on this part to find the most frequent words used in fatwas' topics. We have generated word clouds for fatwas' topics. Figure 2a–m show the word clouds of fatwas' topics for each website. It is clear that the most frequent words are: حكم(Hukm) rule, الله(Allah) Allah, الصلاة(Alsala) the prayer, and المرأة(Almaraah) the woman.

Comparing the word clouds shows that Al Shaikh Abdullah Al Manee's website has a unique word cloud. While the other websites share mostly the same frequent words, Al Shaikh Abdullah Al Manee's fatwas have different words. Most of Al Shaikh Al Manee's fatwas revolve around financial topics. For instance, some of the top frequent words are البنك(Albank) the bank, تمويل(Tamwil) finance, قرض(Qard) loan, بطاقة(Bitaqa) card, and الائتمانية(Aliaytimania) credit. This might be because of the nature of his work and positions. Al Shaikh Al Manee works as a consultant to several Islamic financial institutions around the world. Also, he has publications and a research interest in the field of Islamic banking. Additionally, he has been the Head of the Shariah Committee at Riyadh Bank since 2002.

Another word cloud worth considering is related to Al Shaikh Saleh Bin Humaid's website. It shows that some of its frequent words appeared only in Al Shaikh Bin Humaid's fatwas topics. These unique words are: العلم(Aleilm) the science, طلب(talab) seeking, اليهود(Alyahud) jews, and ضلال(dalal) deception. This might be because of his studies and work in the academic field. Al Shaikh Bin Humaid has a Doctorate of Philosophy in Al-Figh (jurisprudence). Also, he has worked in several positions at Umm al-Qura University: as Teaching Assistant, Lecturer, Assistant Professor, Manager of the Islamic Postgraduate Studies Center, Vice Dean of Al-Shariaa College, and Dean of AlShariaa College.



(**a**) (**b**)

(**c**) (**d**)

(**e**) (**f**)

(**g**) (**h**)

Figure 2. *Cont.*

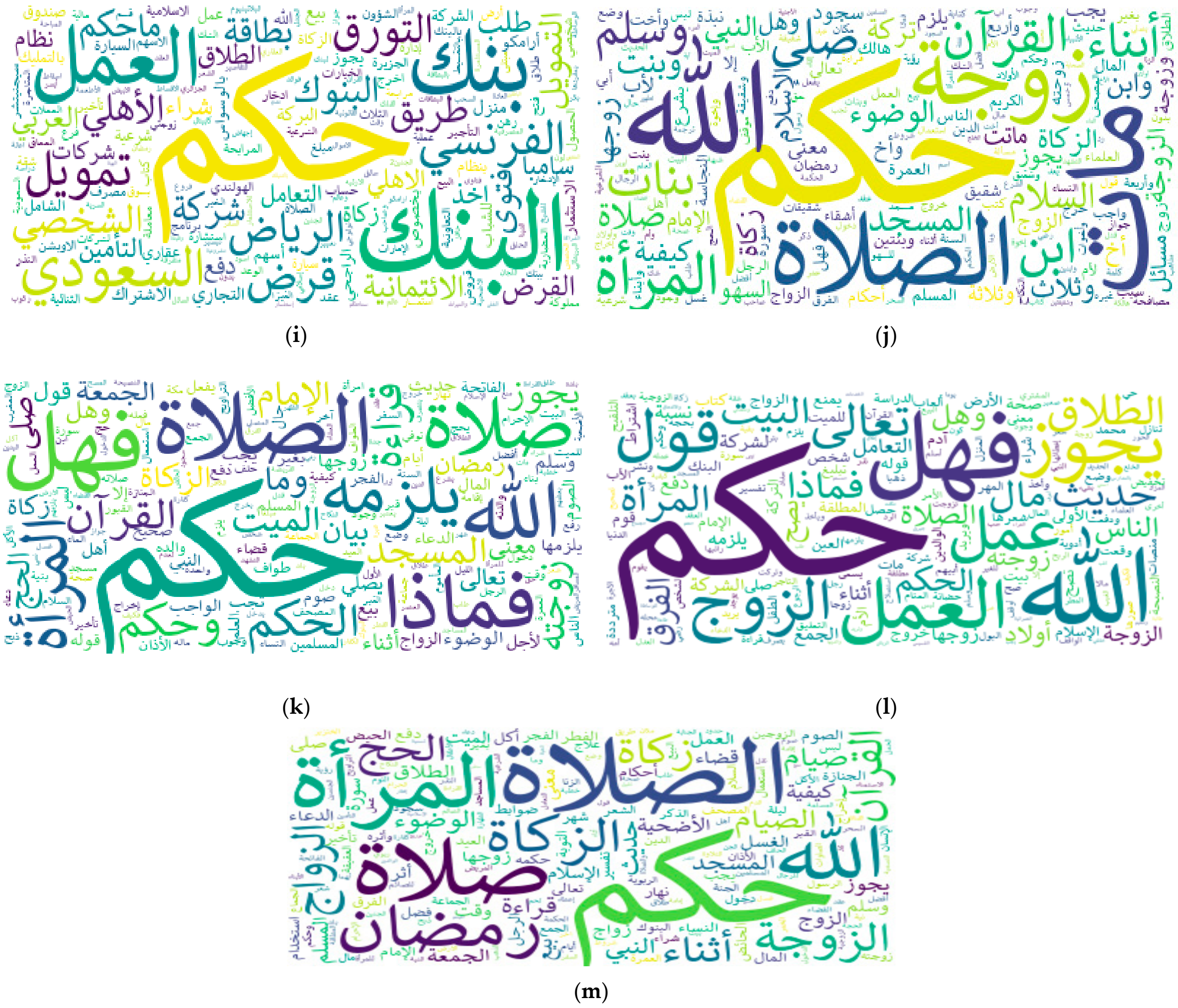


Figure 2. Results of topic modeling: (**a**) Dar Al Ifta in Saudi Arabia; (**b**) Dar Al Ifta in Egypt; (**c**) Dar Al Ifta in Jordan; (**d**) Al Shaikh Abdual Aziz Ibn Baz; (**e**) Al Shaikh Mohammad Ibn Othaimin; (**f**) Al Shaikh Abdual Aziz Al Ashaikh; (**g**) Al Shaikh Saleh Al Fwzan; (**h**) Al Shaikh Saleh Bin Humaid; (**i**) Al Shaikh Abdullah Al Manee; (**j**) IslamWeb; (**k**) FatwaPedia; (**l**) IslamQA; (**m**) IslamOnline.

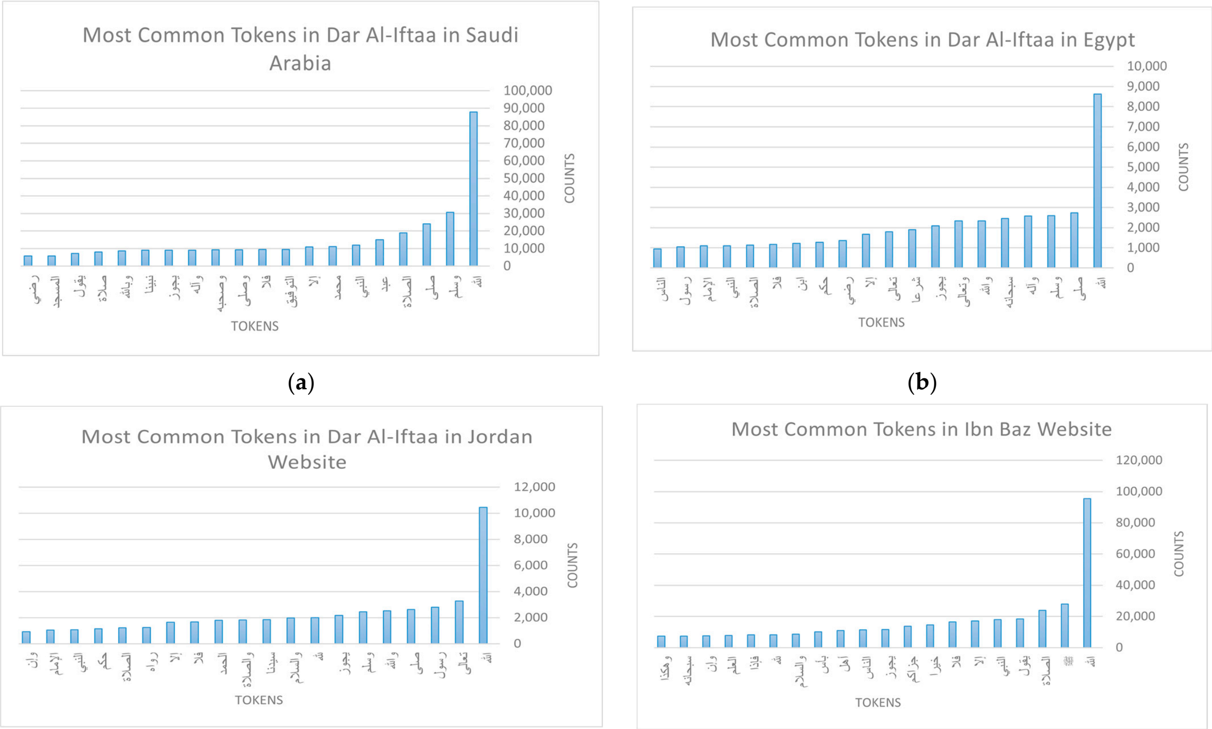
This investigation reveals the topics that are constantly asked about by Muslims and the directions of fatwas. This could be a starting point for further research to discover important facts and provide valuable solutions.

4.2. Fatwas (Questions and Answers)

The other important dimension of this Exploratory Data Analysis (EDA) focuses on the fatwa itself by considering the text body of the fatwa (question and answer). Here, we took each website separately and generated a list of their most frequent words, along with the number of occurrences for each token. Figure 3a–m illustrate the top 20 frequent words regarding fatwas for each website. The words are translated in Appendix A (Table A1) in the order they appear in the figure.

This form of analysis helped us to better understand the data and uncover hidden information. Consequently, it gives insights about crucial details that should be considered when using the data. For instance, the statistics show that most of the websites have the following words as the top frequent words:الحمد(Alhamd) praise, لله(lilah) to Allah. There are also السلام(Alsalam) peace, ورحمة(warahimah) mercy, and وبركاته(wabarakatuh) his blessings. All of these words share a common relationship in which they frequently appear along with the same group of words. For instance, the former set of words is commonly said as a preliminary statement when answering fatwas. The later set of words form the greeting sentence in Islam. In return, this revealed to us that such highly frequent patterns in the dataset may affect the learning mechanism when building a language model, for example, and impact the general use of the data. Thus, it is evident that a specific cleaning should be performed to improve the quality of the data. Accordingly, we provide two versions of the dataset: the original collected dataset and a second version after conducting a cleaning of frequent patterns.

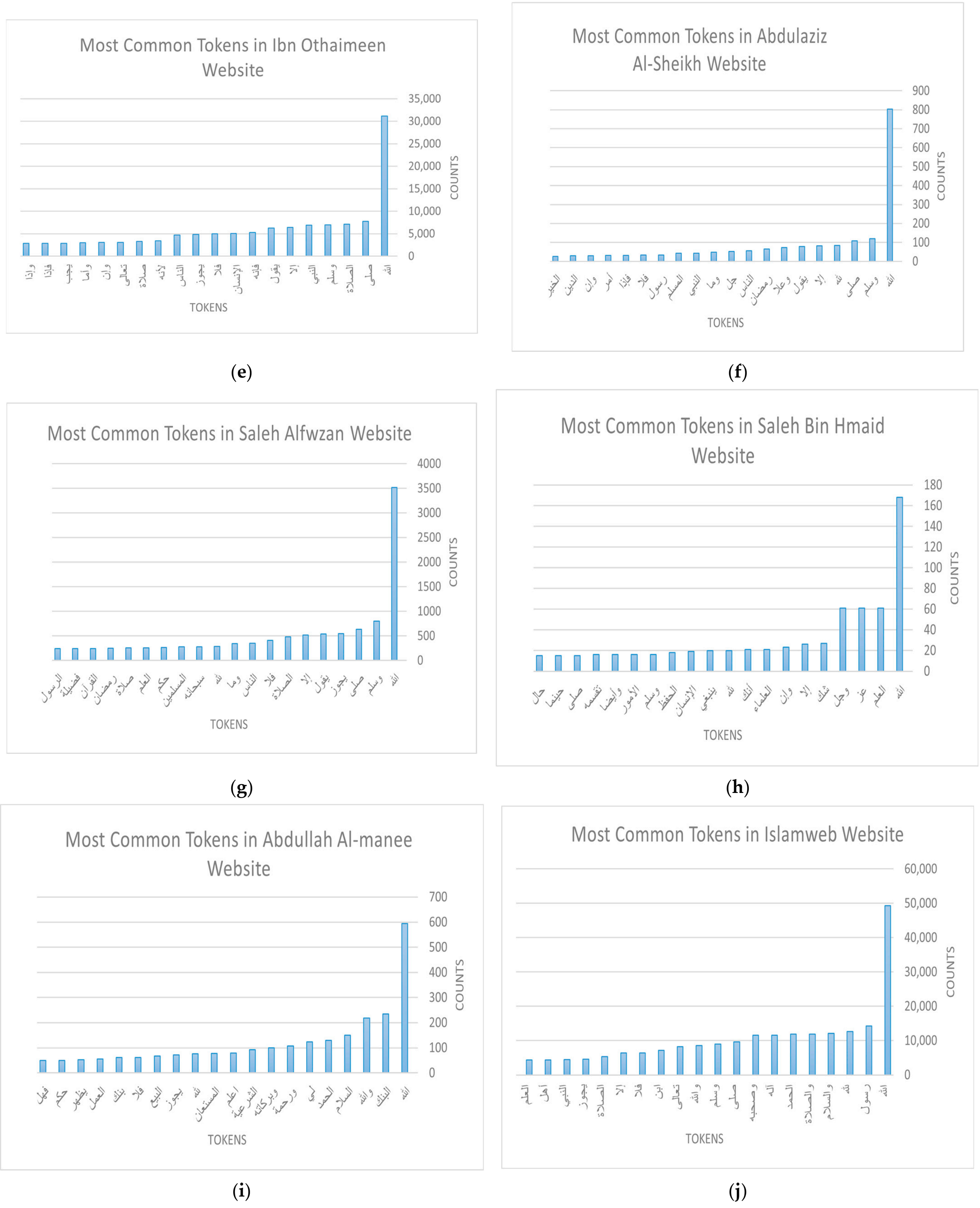
To conduct the desired cleaning, we had to scan the data of each website separately and look for such statements. We could not conduct the same cleaning processes for all websites because such statements take different forms. In other words, each religious scholar has his own way of answering fatwas. For instance, some scholars always start their answers by saying الحمد لله(Alhamd lilah) praise to Allah while others start with the greeting السلام عليكم ورحمة الله وبركاته(alsalam ealaykum warahmat allah wabarakatuh) may the peace, blessings, and mercy of Allah be upon you. Also, some scholars say a specific sentence at the end of the answer, such as والله أعلم(wallah 'aelam) Allah know best. On the other hand, the questions themselves have some frequent patterns that take place at the beginning and at the end of the question text. For instance, the sentence جزاكم الله خيرا(jazakum allah khayran) may Allah reward you. Moreover, the websites add certain lines to questions and answers explicitly such as المقدم: بارك الله فيكم(Almuqadam: barak allah fikum) the interviewer: may Allah bless you. This indicates a sentence that is said by the interviewer. Those sentences required manual checking to find them and a careful approach to remove them. The usual find and remove process is not suitable here because some of these words sometimes occur as part of the body of an answer (where it carries a different meaning). Thus, the context that surrounds a certain set of words should be considered. As a result, we had to use manual cleaning and automated cleaning only in some cases. Such cleaning required months of cleaning and extensive efforts, especially as there are 13 websites and each one has a great number of fatwas and a different style.



(**a**) (**b**)

(**c**) (**d**)

Figure 3. *Cont.*



(e)

(f)

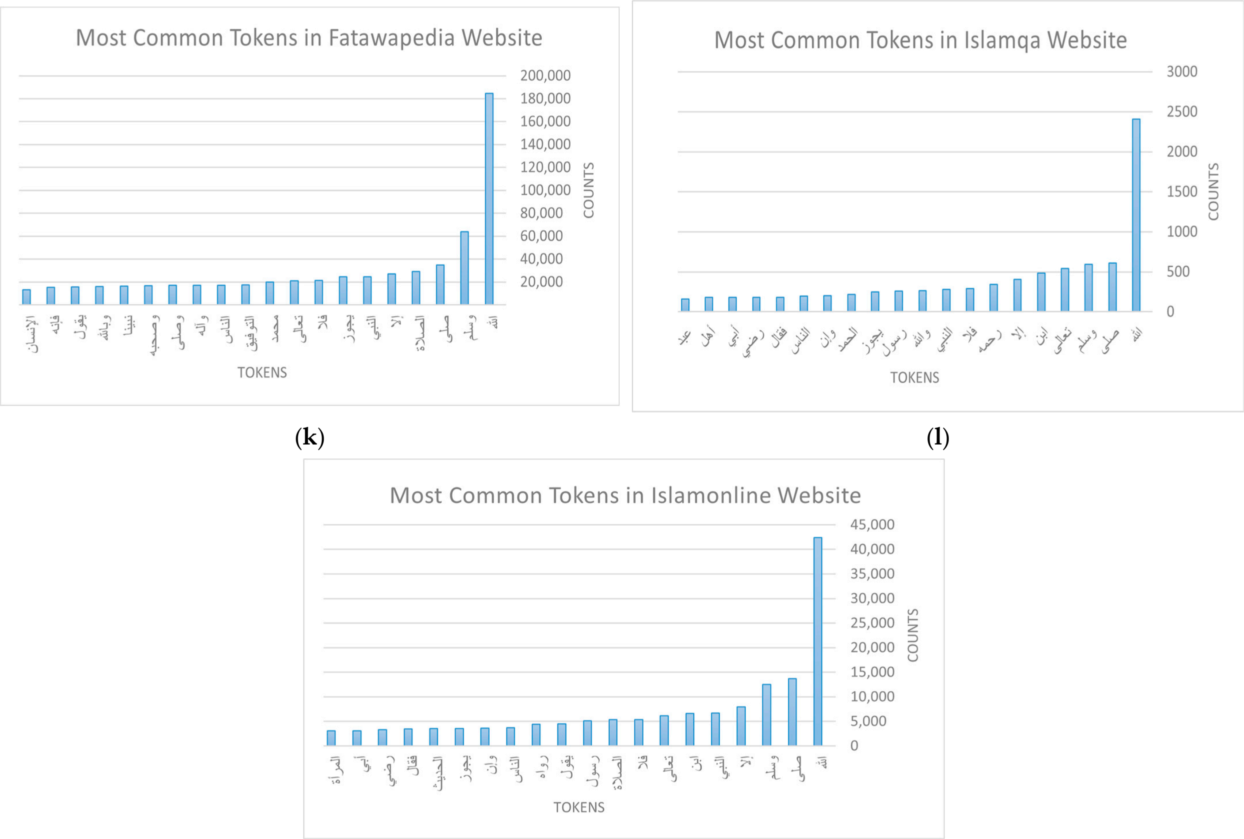
(g)

(h)

(i)

(j)

Figure 3. *Cont.*



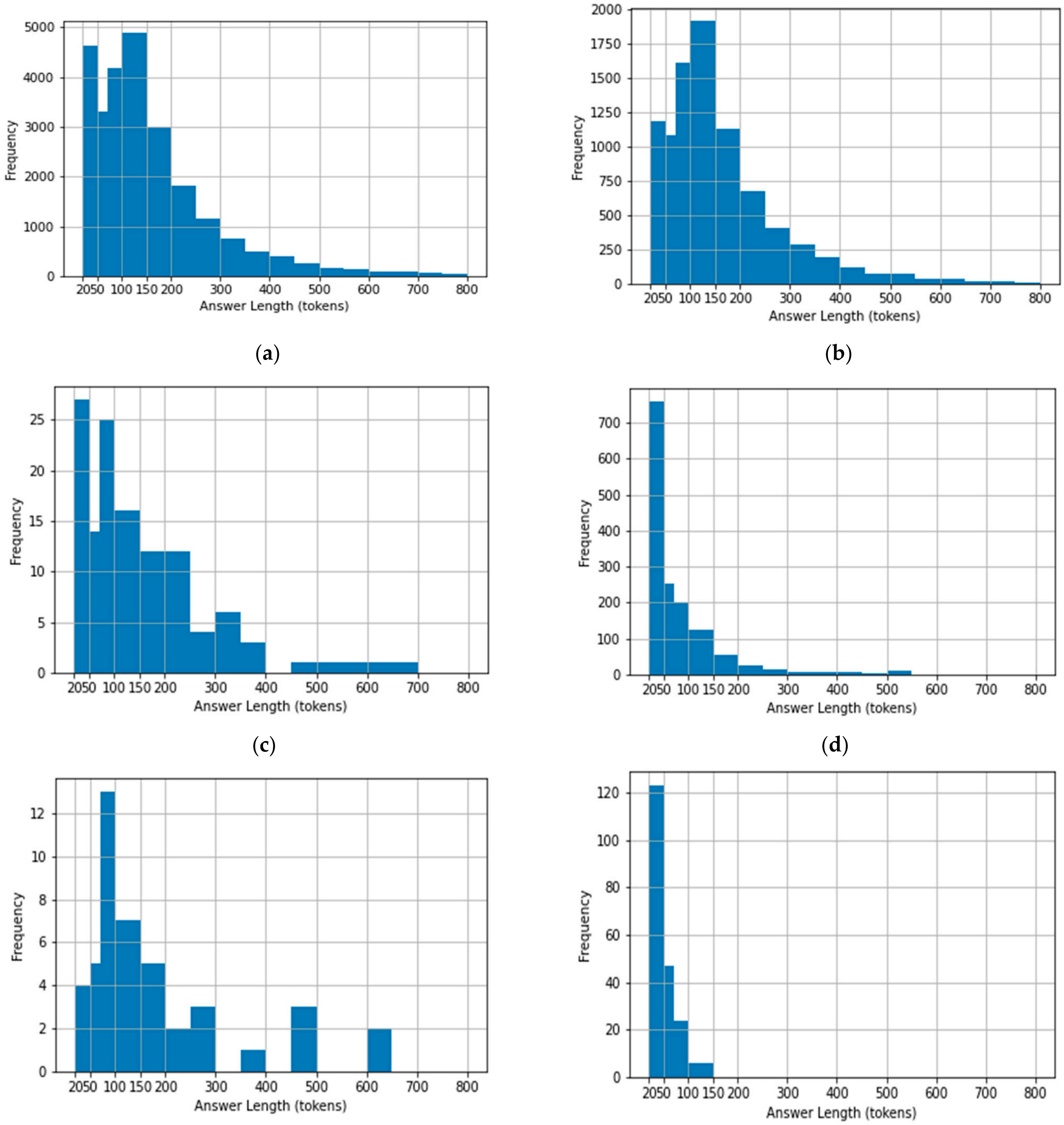
(**k**) (**l**)

(**m**)

Figure 3. Results of EDA (fatwas): (**a**) Dar Al Ifta in Saudi Arabia; (**b**) Dar Al Ifta in Egypt; (**c**) Dar Al Ifta in Jordan; (**d**) Al Shaikh Ibn Baz; (**e**) Al Shaikh Ibn Othaimin; (**f**) Al Shaikh Abdual Aziz Al Ashaikh; (**g**) Al Shaikh Saleh Al Fwzan; (**h**) Al Shaikh Bin Humaid; (**i**) Al Shaikh Al Manee; (**j**) IslamWeb; (**k**) FatwaPedia; (**l**) IslamQA; (**m**) IslamOnline.

4.3. Religious Scholars' Answers

Because the dataset contains fatwas from certain religious scholars' websites, we wanted to examine the way each one of them answers fatwas. Specifically, we plotted histograms that show the number of tokens (words) a scholar used when answering questions. This provides information about the details each scholar gives when responding to fatwas. Also, this may give indications about the complexity of the questions and the personality of each scholar. This step is considered the first step in a research field called stylometry, which is a statistical analysis that studies the linguistic features of a text [19]. We analyzed the answers of all the scholars in the dataset (Al Shaikh Abdual Aziz Ibn Baz, Al Shaikh Mohammad Ibn Othaimin, Al Shaikh Abdual Aziz Al Ashaikh, Al Shaikh Saleh Al Fwzan, Al Shaikh Saleh Bin Humaid, and Al Shaikh Abdullah Al Manee). Figure 4a–f illustrate the histograms of the scholars' answers. The histograms represent the differences in the number of used tokens among the scholars by plotting the frequency of each answer length (in terms of tokens). In general, none of them exceeded 700 words in their answers except Al Shaikh Abdual Aziz Ibn Baz and Al Shaikh Mohammad Ibn Othaimin, who reached 800 words in some of their answers. Al Shaikh Abdullah Al Manee is a special case as he did not exceed 150 words. Regarding the frequency, it is noticeable that Al Shaikh Mohammad Ibn Othaimin and Al Shaikh Saleh Bin Humaid usually gave the longest answers compared with others in the dataset. This is because they frequently answered using, on average, a range of 70-150 words. On the other hand, Al Shaikh Saleh Al Fwzan and Al Shaikh Abdullah Al Mane frequently gave short answers, as their answers are usually in the range of 20–50 words.



(**a**) (**b**)

(**c**) (**d**)

(**e**) (**f**)

Figure 4. Histograms of religious scholars' answers: (**a**) Al Shaikh Ibn Baz; (**b**) Al Shaikh Ibn Othaimin; (**c**) Al Shaikh Abdual Aziz Al Ashaikh; (**d**) Al Shaikh Saleh Al Fwzan; (**e**) Al Shaikh Bin Humaid; (**f**) Al Shaikh Al Manee.

5. Conclusions

In this paper, we have reviewed the literature on Islamic content and highlighted important points for interested researchers. We have also introduced an Arabic dataset of Islamic fatwas, *fatwaset*, that is available to the public. *Fatwaset* is designed to address the limitations of Islamic content datasets. *Fatwaset* can be integrated into various Natural Language Processing (NLP) systems for the purposes of question answering, text classification, text generation, or author attribution. Also, researchers from other fields, such as art and philosophy, can use *fatwaset* to analyze and study its Islamic content from different perspectives. Additionally, we performed an Exploratory Data Analysis (EDA) on the dataset. *Fatwaset* and the results of the EDA contribute to understanding Islamic content as well as to the development of new tools and methods for processing and analyzing Islamic content. It is worth mentioning that *Fatwaset* has been collected from a subset of the available Islamic fatwas' websites in some Arabic countries. Also, the dataset contains only fatwas in text format. The number of fatwas collected from each website is not the same because some websites contain more fatwas than other websites in the dataset.

Author Contributions: Conceptualization, O.A. and H.A.-K.; methodology, O.A. and H.A.-K.; software, O.A.; validation, O.A. and H.A.-K.; formal analysis, O.A. and H.A.-K.; investigation, O.A. and H.A.-K.; resources, O.A. and H.A.-K.; data curation, O.A.; writing—original draft preparation, O.A.; writing—review and editing, H.A.-K. and A.M.; visualization, O.A., H.A.-K. and A.M.; supervision, A.M. and H.A.-K. All authors have read and agreed to the published version of the manuscript.

Funding: The authors would like to thank Deanship of scientific research in King Saud University for funding and supporting this research through the initiative of DSR Graduate Students Research Support (GSR).

Informed Consent Statement: Not applicable.

Data Availability Statement: Data can be found here: https://github.com/ohoud/Fatwaset.git (accessed on 18 October 2023). The data on each website are saved in a separate excel file.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Translation of the tokens presented in Figure 3.

Token (In Arabic)	Translation	Token (In Arabic)	Translation
الله	Allah (God)	العلم	the science
وسلم	and peace be upon	وهكذا	and so on
صلى	prayed	الإنسان	the human
الصلاة	the prayer	يقول	says
عبد	slave	وعلا	exalted
النبي	The Prophet	رمضان	Ramadhan
محمد	Mohammad	جل	Majestic
إلا	unless	المسلم	The Muslim
التوفيق	success	الدين	The religion
فلا	and no	الخير	The good
وصحبه	his companions	المسلمين	The Muslims
واله	his family	القرآن	The Quraan
يجوز	Permissible	فضيلة	virtue
نبينا	Our prophet	عز	Almighty
وبالله	and with Allah (God)	شك	doubt
صلاة	prayer	العلماء	The scientists
يقول	said	ينبغي	should
المسجد	The mosque	الحفظ	Preservation
سبحانه	Glorified	الأمور	matters
وتعالى	exalted	وأيضا	additionally
شرعا	legally	تقسمه	divide it

Table A1. *Cont.*

Token (In Arabic)	Translation	Token (In Arabic)	Translation
حكم	rule	حينما	when
ابن	son	حال	status
الإمام	Imam (leader)	البنك	the bank
رسول	Messenger	وبركاته	His blessings
الناس	the people	الشرعية	legitimacy
سيدنا	Our master	أعلم	Know best
الحمد	praise	المستعان	The helper
رواه	Narrated by	البيع	The sale
وإن	and that	العمل	The work
خيرا	good	يظهر	shows
جزاكم	reward you	أبي	Father of
أهل	people	الحديث	Hadith
فإذا	and if	المرأة	The woman

References

1. Munshi, A.A.; Al-Sabban, W.H.; Farag, A.T.; Rakha, O.E.; Alotaibi, M.; Alotaibi, M. Towards an Automated Islamic Fatwa System: Survey, Dataset and Benchmarks. *Int. J. Comput. Sci. Mob. Comput. (IJCSMC)* **2021**, *10*, 118–131. [CrossRef]
2. Al-Yahya, M. Towards Automated Fiqh School Authorship Attribution. In *Computational Linguistics and Intelligent Text Processing*; Gelbukh, A., Ed.; Springer International Publishing: Cham, Switzerland, 2018.
3. Abdullah, O.; Shaharuddin, A.; Wahid, M.A.; Harun, M.S. The Potential and Challenges of Decision Support Systems for Islamic Banking and Finance. *Eur. J. Islam. Financ.* **2022**, *9*, 21–29. [CrossRef]
4. Khairuldin, W.M.K.F.; Anas, W.N.I.W.N.; Embong, A.H.; Hassan, S.A.; Hanapi, M.S.; Ismail, D. Ethics of Mufti in the Declaration of Fatwa According to Islam. *J. Leg. Ethical Regul. Issues* **2019**, *22*.
5. Alnefaie, S.; Atwell, E.; Alsalka, M.A. Challenges in the Islamic Question Answering Corpora. *Int. J. Islam. Appl. Comput. Sci. Technol.* **2022**, *10*, 1–10.
6. Malhas, R.; Elsayed, T. AyaTEC: Building a Reusable Verse-Based Test Collection for Arabic Question Answering on the Holy Qur'an. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **2020**, *19*, 1–21. [CrossRef]
7. Malhas, R.; Mansour, W.; Elsayed, T. Qur'an QA 2022: Overview of The First Shared Task on Question Answering over the Holy Qur'an. In Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection; OSACT; European Language Resources Association: Marseille, France, June 2022; pp. 79–87.
8. Mohammed, M.; Amin, S.; Aref, M.M. An English Islamic Articles Dataset (EIAD) for developing an IslamBot Question Answering Chatbot. In Proceedings of the 2022 5th International Conference on Computing and Informatics (ICCI), Riyadh, Saudi Arabia, 9–10 March 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 303–309. [CrossRef]
9. AlZahrani, F.M.; Al-Yahya, M. A Transformer-Based Approach to Authorship Attribution in Classical Arabic Texts. *Appl. Sci.* **2023**, *13*, 7255. [CrossRef]
10. Gartner, R. *Metadata: Shaping Knowledge from Antiquity to the Semantic Web*; Springer: Cham, Switzerland, 2016; pp. 1–10.
11. Riley, J. *Understanding Metadata: What Is Metadata, and What Is It For?: A Primer*; NISO: Baltimore, MD, USA, 2017; pp. 1–49.
12. Sahoo, K.; Samal, A.K.; Pramanik, J.; Pani, S.K. Exploratory Data Analysis Using Python. *IJITEE* **2019**, *8*, 4727–4735. [CrossRef]
13. Endsuy, R.D. Sentiment Analysis between VADER and EDA for the US Presidential Election 2020 on Twitter Datasets. *JADS* **2021**, *2*, 8–18. [CrossRef]
14. Komorowski, M.; Marshall, D.C.; Salciccioli, J.D.; Crutain, Y. Exploratory Data Analysis. In *Secondary Analysis of Electronic Health Records*; Springer: Cham, Switzerland, 2016; pp. 185–203. [CrossRef]
15. Kalmukov, Y. Using Word Clouds for Fast Identification of Papers' Subject Domain and Reviewers' Competences. *arXiv* **2021**, arXiv:2112.14861. [CrossRef]
16. Balz, T. Scientometric Full-Text Analysis of Papers Published in Remote Sensing between 2009 and 2021. *Remote Sens.* **2022**, *14*, 4285. [CrossRef]
17. Alfraidi, T.; Abdeen, M.A.; Yatimi, A.; Alluhaibi, R.; Al-Thubaity, A. The Saudi Novel Corpus: Design and Compilation. *Appl. Sci.* **2022**, *12*, 6648. [CrossRef]

18. Albadi, N.; Kurdi, M.; Mishra, S. Are they our brothers? Analysis and detection of religious hate speech in the Arabic twittersphere. In Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Barcelona, Spain, 28–31 August 2018; IEEE: Piscataway, NJ, USA; pp. 69–76. [CrossRef]
19. Adebayo, G.O.; Yampolskiy, R.V. Estimating intelligence quotient using stylometry and machine learning techniques: A review. *Big Data Min. Anal.* **2022**, *5*, 163–191. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.