A Multilingual Datasets Repository of the Hadith Content

Ahsan Mahmood
Department of Computer Science
COMSATS Institute of information Technology,
Attock, Pakistan

Fawaz K. Alarfaj
Computer and Information Science Department
Al-Imam Mohammad Ibn Saud Islamic University, Al-Hofuf, Kingdom of Saudi Arabia

Hikmat Ullah Khan*
Department of Computer Science
COMSATS Institute of information Technology,
Wah, Pakistan

Muhammad Ramzan, Mahwish Ilyas
Department of Computer Science and Information Technology, University of Sargodha,
Sargodha, Pakistan

***Abstract*—Knowledge extraction from unstructured data is a challenging research problem in research domain of Natural Language Processing (NLP). It requires complex NLP tasks like entity extraction and Information Extraction (IE), but one of the most challenging tasks is to extract all the required entities of data in the form of structured format so that data analysis can be applied. Our focus is to explain how the data is extracted in the form of datasets or conventional database so that further text and data analysis can be carried out. This paper presents a framework for Hadith data extraction from the Hadith authentic sources. Hadith is the collection of sayings of Holy Prophet Muhammad, who is the last holy prophet according to Islamic teachings. This paper discusses the preparation of the dataset repository and highlights issues in the relevant research domain. The research problem and their solutions of data extraction, pre-processing and data analysis are elaborated. The results have been evaluated using the standard performance evaluation measures. The dataset is available in multiple languages, multiple formats and is available free of cost for research purposes.**

Keywords*—Data extraction; preprocessing; regex; Hadith; text analysis; parsing

I. Introduction

Data mining, Information retrieval and knowledge extraction have become attractive fields for the researchers during the last decade due to the birth of Social media [1]. These fields are getting researchers' interest because textual data over the internet is expanding exponentially during the last decade. The internet users are shifting from conventional methods of communications to online social networks at a rapid rate [2]. Sharing textual data over the internet is common due to the Social media channels [3]. Data mining and knowledge discovery tasks are carried out using the machine learning, statistical and database oriented approaches [4]. For the purpose of Knowledge discovery, researchers have used databases of different languages and domains to meet the requirements [5]. The recent research focuses on diverse techniques to make the unstructured data over the internet to be converted into such structured form so that various text mining and content analysis tasks can be accomplished and the data become more understandable as well machine readable [6].

Hadiths are regarded as one of the major sources of knowledge of the religion of Islam. The Hadith are the sayings of the Holy Prophet Muhammad, who is the last apostle according to Muslims. Analyzing Hadith text results in knowledge discovery from Hadith with the help of natural language processing methods. Although many researchers work in this field, there is no work solely focused on Hadith data. Moreover, it is not possible to compare these works with one another due to the unavailability of the common Hadith data corpus. There are a number of web sources which contains the Hadith contents. However, there is a lack of a repository containing data sets of Hadith for researchers to work in various research domains, such as text mining, data analysis, information retrieval and knowledge extraction. In this paper, we focus on preparation of a repository of the Hadith content data sets. The Hadith content is extracted from the reliable online sources. The volume of the data related to Islamic knowledge is present in a huge amount and is available in two major forms including the Quran and Hadith. Many researchers around the world who have worked on Natural language processing tasks, used Quran and Hadith data for knowledge discovery and Data mining tasks. However, most of the times, researchers have used Quran data for their research and knowledge discovery and have overlooked Hadith data. One of the most important reasons is the unavailability of data corpus of Hadith data [7]. A number of data mining researchers develop their own data corpus. Some of the researchers used the existing datasets, but those datasets are not present in enough amount considering the real data of Hadith. After a Data corpus of Hadith become available, it will become easy for the researchers to achieve Data mining and knowledge discovery tasks on Hadith data and compare performance of different works in the field.

In this research paper, we discuss our research contribution for the Hadith data repository preparation from different websites, processing them through different techniques and preparation of a data set repository of Hadith content that can

further be used for knowledge discovery and Data mining tasks by researchers. The rest of the paper is as follows: Section II reviews earlier studies, Section III discusses online Hadith resources used, Section IV discusses the details of the proposed research methodology and Section V discusses experimental setup and evaluation of our results before concluding the paper.

II. Background

Muslims believe that Muhammad (Peace and Blessings may Allah be upon him (PBUH)) is the last messenger of Allah. In religious terms, Hadith, meaning "tradition", is a report of the actions and sayings of Prophet Muhammad (PBUH). There are a number of Hadith books, but mainly six books, known as Sihah-e-Sita are regarded as the most authentic books.[1] The six most authentic books are Sahih Bukhari, Sahih Muslim, Sunnah Abu Dawood, Sunnah Nasai, Sunnah Tirmidhi, and Sunnah ibn Majah. After the Holy book of Quran, Hadiths are regarded as the second most important source of guidance in the Islam. Each Hadith consists of two things, the chain of narrators called isnad, and Hadith text called meeting. Rawi AL-Hadith, the person who reports a prophetic tradition is called Narrator or Rawi of Hadith. In all the books of Hadith, the content is divided into multiple parts. Usually, a book consists of volumes, each volume contains multiple chapters and each chapter has many Hadith referred in it. Each Volume and chapter has its own name and number while each hadith is assigned a number, list of narrator(s) and its content. While there are many parts of Hadith, the most important parts of Hadith are sanad and Matn that contains the actual textual content. Fig. 1 presents the hierarchy of a Hadith content in Hadith books.

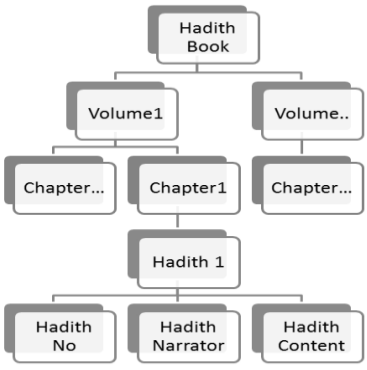


Fig. 1. Structure of the Hadith Data

A Hadith book consists of many volumes. Each volume consists of a number of chapters. The chapters are the topic-wise pre-categorization of Hadith. For data extraction, we extract all the required attributes that are associated with each Hadith including volume number and name, chapter number and name, hadith number, Sanad and Hadith matn.

III. Related Work

In the modern era, the amount of Information available on the web is growing at a very fast pace and have become the largest source of knowledge for the users, however processing and accessing such a large amount of data is not possible due to its vastness. [8]. Although the web traffic is constantly increasing thus the usage of regular expression is also increasing for packet inspections. Usually Regex matching is slow due to backtracking, but some new work has been done in this field using future matching techniques and achieve up to 70% performance boost [9]. Although many researchers have used the Hadith data for research purposes, there is no combined research work done in this regards. None of those researchers are able to compare their work with the others because there is no proper dataset of Hadith. In different hadith books, data has been divided into multiple parts and each book has its own format so each book takes its own time to extract data. Moreover, Hadith data is available in different languages that make it difficult to perform and compare the results of natural language processing tasks or to propose a framework of data extraction of Hadith that can be used with all the books or languages. Many researchers have worked to extract the hadith data from separate books using different kinds of techniques.

Aldhlan et al. [10], presents their understanding of new techniques in data mining to collect the Islamic knowledge from multiple resources, and represent the knowledge to the users in a better way. In their work, they used Hadith as a source of knowledge and proposed an approach for the classification of Hadiths in multiple categories using Supervised learning. They also discussed several ways to extract the knowledge from Hadith related to the goal of knowledge. Harrag et al., [11] perform Named Entity Recognition, extraction of words and entities from Hadith data. Siddiqui et al., [12] Proposed a system to extract isnad from Hadiths using Named entity extraction and classification in a form of network. Another work discusses extraction of the surface information from Sahih Bukhari. Author proposed a system based on Finite State Transducers (FST) to extract the knowledge from hadith text and their work shows 71% and 39% precision and recall respectively [13]. Another work proposed for NER for Hadith documents using Maximum Entropy Classifier, Naïve Bayes Classifier and Support Vector Machine based techniques and achieved an F-measure of 95.3% [14]. Mahmood A et al [15], proposed a framework for knowledge extraction from Sahih Bukhari urdu translation book. Other than named entity extraction, different sorts of other techniques have been used to extract the data from Hadith books. The method of handling missing data while extracting the Hadith data has been presented which is called missing data detector (MDD) [16]. The authors proposed a method based on the Isnad validity methods in Hadith science. Another tool proposed on the basis of Vector Space Model to let the users search for a particular Hadith from the complete Data repository with better precision. [17]. The authors performed Hadith classification according to the similarity between them.

The extraction of chain of narrators is also an important step. The authenticity of Hadith depends on authentication of the chain of narrators' and its content. The first step in the hadith authentication process is to determine chain of narrators' authentication. The chain of narrators' can be represented in the form of network graph. Network graph has

[1] https://en.wikipedia.org/wiki/Hadith. Accessed on August 14, 2016.

an element of the chain that can help in search of chain of narrators' for hadith [18]. The authentication measurement of the Hadith data is also important. Another research work share a reliable method to extract Hadith text from Islamic web pages [19]. In this work, researchers used Shiekh Al-Albani Hadith Database collection and finds out the correctness of each item. Working with different languages of Hadith has different experience due to their structures and diversity in the format of Hadith in each book. Therefore, it is not possible to propose a single algorithm for data extraction from the Hadith that can be applied to all the Hadith books of different languages. A study examined the knowledge discovery from Al-Hadith content by using classification algorithms [20]. It classifies the hadith content into one of the pre-defined books of hadith (classes in terms of classification). According to them Arabic language has a complex morphological structure and orthography variations. An android based application [21] targets Hadith retrieval system. Arabic is a major language in around two dozen countries of the world [22]. It differs from conventional retrieval systems because it allows Hadith retrieval in non-conventional manner. It allows the users to search for Hadiths using root based search. These kinds of sophisticated searches require extensive database, however, we kept the database as simple as possible and utilized regular expression (RE), which is supported by many modern programming languages.

According to another model, which create parts of each unit in *isnad* and *matn* and further process each part. It also creates a graph based on the relation between transmitters using an AraMorph morphological analyzer (RAM) and explains the text content [23]. In addition to hadith data, the work done on Quranic content is also valuable to mention. In the holy Quran the semantic web ontology has been applied for the purpose of search and extraction of semantic knowledge, including Quranic Wordnet, and mapping of domain ontology with higher level ontologies and it can also be applied to Hadith Data. [24]. Different kinds of models have been presented to perform the semantic search in the Holy Quran. A relational WordNet model [25] is presented to perform the semantic search in the Holy Quran that has been carried out in the latest tools and researchers used Surah AlBaqrah as a sample and produced their results on that basis. Quranic Arabic Dependency Treebank (QADT) model [26] reports on the approaches and solutions used in applying NLP to the Challenging Language of Quran. The authors proposed a complex linguistic model based on the Arabic language. It has been argued that memorization and methodologies are important factors that enhance the practices of memorization in the Islamic world [27].

Alqahtani M et al. [28], discussed different search techniques and proposed a model built on those techniques on Quran for searching purposes, including ontologies and semantic search tools for holy Quran. A sub-path mining algorithm built for the Holy Quran content to generate frequent patterns that can also be used for indexes and clusters in Quran Data [29].

Hadith books are not present in the form of Dataset on the internet. All the researchers who work on Hadith data develop their own Hadith dataset and do their experiments. In this research our focus is to collect Hadith data across different websites and develop a central Data corpus where users can download any Hadith book dataset. For this purpose, we use different websites to collect data of Hadith. The source websites are discussed in the next section. During data crawling, we faced some problems like slowness and noise in the web data, but we retrieved data chapter by chapter and volume by volume so data can be easily manageable. For Sahih Muslim data, we used SahihMuslim.com website that has all the Hadith of Sahih Muslim in a better format that can easily be crawled. There are some other websites that we have used for data crawling purpose as mentioned at the start.

IV. Research Methodology

In this research work, we select a number of sources and process textual data of Hadith and develop datasets of different books. Regular expressions [30] are used for data extraction purpose. We discuss the details about our extraction process that we use to extract the Hadith data from different sources.

A. Selection of Hadith Sources

The authentic and reliable sources are selected from the sources of Hadith content. Data from the sources, whether in the form of text on websites or in the form of documents such as in PDF form have been used for data extraction. Although Hadith books are present over the internet in a number of formats and types, we focus on Hadith books available in Unicode format so the data present in the book can be easily processed. The sources from which data is taken are as follows:

1) Hadith Websites

There are a number of websites that contain reliable Hadith content, possesses such a structure which allows us to apply regex and retrieve the desired data by matching the patterns [31]. It is notable that websites with AJAX [32] (asynchronous JavaScript and XML) do not allow its users to scrap the website content. Table I shows the list of books downloaded for extraction along with their source.

TABLE I. Hadith Book Sources

S#	Book Name	Source
1	Sahih Muslim English	http://sunnah.com/muslim
2	Sunan Abu Dawud	http://ahadith.co.uk/
3	Mawta Imam Malik	http://ahadith.co.uk/maliksmuwatta
4	Sahih Al-Bukhari	http://www.sahih-bukhari.com/, http://hadithcollection.com/

B. Hadith Content Extraction

Regular expressions search for a particular pattern and retrieve output based on the pattern. It helps to extract the required tokens from Hadith Data but we face the issues of data variability as each book has its own format and in each book there are variations of length, content and difference in structure.

In case of Sahih Bukhari Data extraction, the proposed regex extracts all the parts of Hadith easily as Sahih Bukhari data on the website is properly managed. Fig. 2 shows the extraction process from Hadith in Sahih Bukhari. Moreover, the entity relationship description diagram (ERD) presented in Fig. 3 that shows the entities as well as the attributes extracted. The diagram shows the structure of Mawta Imam Malik Hadith book. Each book has its own structure and we need to create a separate database design for each book. Fig. 3 shows an ERD diagram of Hadith Book "Mawta Imam Malik".

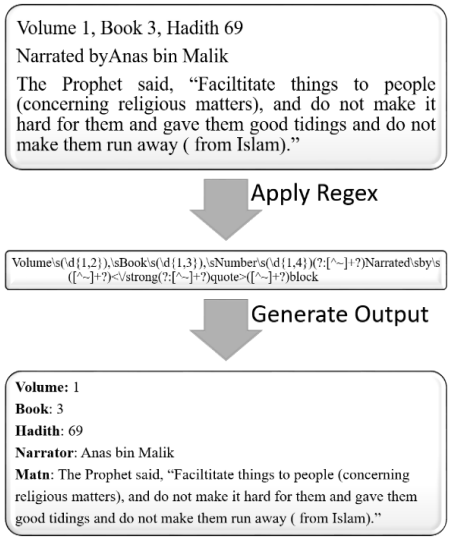


Fig. 2. Sahih Bukhari data extraction process.

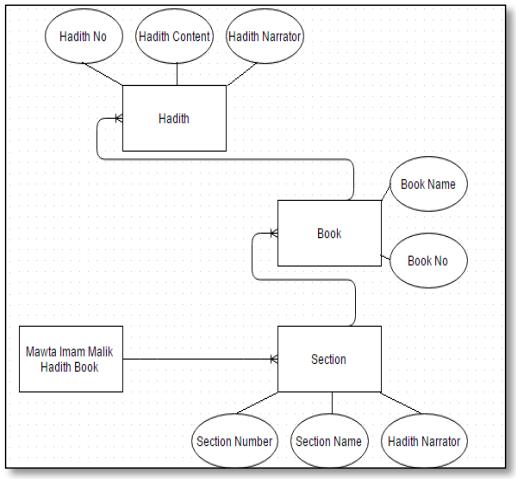


Fig. 3. ERD diagram for database of Mawta Imam Malika Hadith book.

1) Regular Expression

Regular expressions or Regex is used to search and retrieve a particular pattern in a stream of text document. The output generated by Regex can further be used to explore the extracted text. According to the context of the textual data, different types of regular expressions used in this research. Writing a regex is a very sensitive work as for detecting a particular text pattern, different regex can be used. The major priority during regex proposition is to write a regex with lowest possible steps and least backtracking. The higher number of steps involves matching a pattern, the higher system resources such as memory and processing power use of the system. According to Fang Yu et al. [33], firm and efficient ways of regex matching are required.

a) Regex example for recognizing Hadith Sanad and Matn

Below is an example of Hadith text. In order to extract the Sanad and Matn from below example a regex can be proposed. Regex shown in Fig. 4 extracts Sanad and Matn part of the hadith from below example.

Narrated by Al-Hasan: 'Amr bin Taghlib said, "Some property was given to the Prophet and he gave it to some people and withheld it from some others. Then he came to know that they (the latter) were dissatisfied. So the Prophet said, 'I give to one man and leave (do not give) another, and the one to whom I do not give is dearer to me than the one to whom I give. I give to some people because of the impatience and discontent present in their hearts, and leave other people because of the content and goodness Allah has bestowed on them, and one of them is 'Amr bin Taghlib." 'Amr bin Taghlib said, "The sentence which Allah's Apostle said in my favor is dearer to me than the possession of nice red camels."

```
Narrated\s(?:[Bb]y|[Ff]rom)\s(.+):\s?([^+]+)(?<=').+?bin.+?(?=\s)
```

Fig. 4. Regex Example3 for extraction of narrator and content.

b) Regex example for recognizing Hadith Number, Sanad and Matn data from Html Source

In the below example, there are three attributes of Hadith that can be extracted. These attributes are Hadith No, Hadith Sanad and Matn. Regex shown in Fig. 5 can be used to Extract those attributes from the below Hadith text.

<aname=18.1.1> 18.1.1

</td><td class="QuranData" bgcolor="#FFFFFF" valign="top"> Yahya related to me from Malik from Nafi from Abdullah ibn Umar that the Messenger of Allah, may Allah bless him and grant him peace, once mentioned Ramadan and said, "Do not begin the fast until you see the new moon, and do not break the fast (at the end of Ramadan) until you see it. If the new moon is obscured from you, then work out (when it should be)."

</td></tr><td class="QuranData" bgcolor="#FFFFFF" valign="top">

```
(\d{1,3}\.\d{1,3}\.\d{1,3}).+(?=;);(.+)that([^+]+(?=&))
```

Fig. 5. Regex Exanmple 4 for extraction of Narrator and Hadith cContent from Html document.

1) Issues and solutions

While working on text processing some minor issues that arises during the data retrieving and saving to database. At some places, those issues are negligible, but some issues are non-negligible. One of the major issues in Hadith text processing is escape sequences in the textual data when saving or retrieving the data into the databases. These escape

sequences become even more challenging when processing the data that is present on a website. In these cases, we remove escape sequence before retrieving. Therefore, sometimes when we process the Hadith data, we are not able to remove the escape sequence from textual data and backslashes are visible in the text when we retrieve the Hadith text for further processing.

C. Dataset Preparation

After text extraction, we execute other steps to in text processing to save the data into the database and develop a proper data corpus. Data preparation steps are shown in Fig. 6.

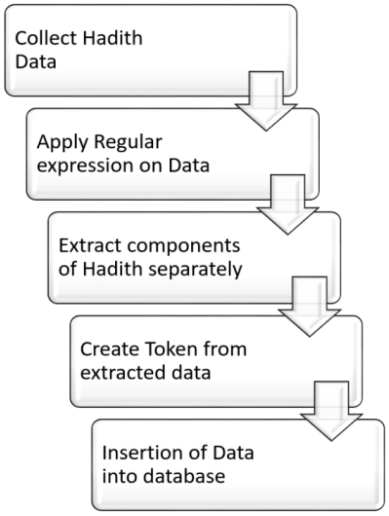


Fig. 6. Dataset preparation steps.

1) Diversity in Dataset preparation

As each of the Hadith books has its own format, we cannot combine all the datasets in one dataset. During dataset preparation and tokenization of the text we faced different kinds of problems.

An example of a webpage is given in Fig. 7. This example shows some Hadiths from Sahih Bukhari website.

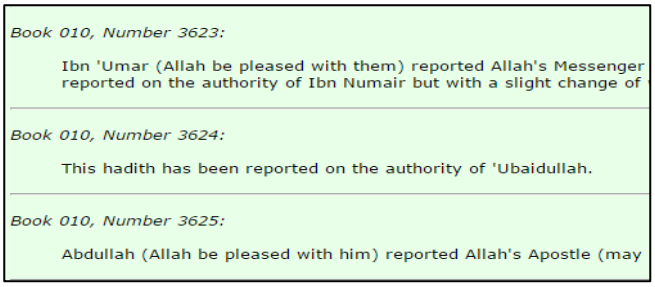


Fig. 7. Hadith website data example.

The problem however with this book is that it does look so simple in the webpage, but a lot of customized fonts etc. are present. Therefore, its source file has a lot of unnecessary things so to cope with these issues our regex became complex. In Fig. 8, it can be seen that the source code has a very huge amount of unnecessary text and small amount of information so proposing and debugging regex in such cases are challenging.

```
14"><strong><img src"..="" image="" book.gif"="" align="absbottom"> Volume 1,
ed by Ibn Umar</strong></td></tr>
same as above Hadith 59. </blockquote></td></tr>
14"><strong><img src"..="" image="" book.gif"="" align="absbottom"> Volume 1,
ed by Ibn Umar</strong></td></tr>
same as above Hadith 59. </blockquote></td></tr>
14"><strong><img src"..="" image="" book.gif"="" align="absbottom"> Volume 1,
ed by Ibn Umar</strong></td></tr>
same as above Hadith 59. </blockquote></td></tr>
14"><strong><img src"..="" image="" book.gif"="" align="absbottom"> Volume 1,
ed by Anas bin Malik</strong></td></tr>
While we were sitting with the Prophet in the mosque, a man came riding on a
was sitting amongst us (his companions) leaning on his arm. We replied, "This
```

Fig. 8. HTML source of Hadith website data.

In these cases, our regex become complex while in case our data is simply available on a webpage with small amount of noise in the data, the regex become simple and it can extract the hadith text with the smallest number of steps required to extract the Hadith text. Some of our proposed regex are given in Table II.

TABLE II. SAMPLE REGEX FOR DIFFERENT BOOKS

Type	Regular Expression
Sahih Bukhari Arabic	(?+[`^])حَدَّثَنَاReference[^`]+?Vol. (\d{1 ,3}),Book (\d), Hadith (\d{1,3})
Sahih Bukhari English	(\d{1,3}),\sBook\s(\d{1,3}),\sNumber \s(\d{1,4})(?::|;)?,?\s\r\n?([^%]+?)(?::| ;|,)?\s?\r\n?([^%]+?)(?:Volume|Book)
Sahih Muslim English Initial	(?:[^%]+?(?:by|of|narratedon)?([^%]+ ?)(?:that|</font))
Sahih Muslim English Final	(?:(\d{1,4}),(?:N|n)umber\s(\d{1,5}) :.?\r\n)(?:(.+?)(?:said|that|narrates|rep orted:|who|that|by.+?\)|authorityin||sa me.+?\r\n|like|by another chainof|\, while|Apostle|somanyahadith|Messg er|addition|reported,|authority.+?\r\n. +?authority|chai|slight|andhe|change ofwords))([^%]+?)(?:Book\s\d|Capp)
Sunan Abudawud English	(?:\s(\d{1,3}),\sNumber\s?(\d{1,4})[^%]+?Narrated\s(.+?):\s?\r\n([^%] +?)Book)
Sunan Abudawud Urdu	\:حدیث نمبرs(\d{1,3})[^%]+?\.\. (([^%]+?)(?:کہتے|روایت|مروی \?+[%^](:کہا|کہتی|r\n\r\n)

Due to the difference in structure of the Hadith books, we use different kinds of Regex for data extraction. In some books data was present in a particular format so extraction was easy while in other books the data was not available in easy format so Regex become complex. In some books, it is not easy to extract Sanad and Matn separately so proposed Regex are very complex thus show. Table II shows some examples of Regex for different books.

D. Website Creation

Fig. 9 shows this, we plan to launch this webpage in the near future; however, dataset a screenshot of the webpage that we have developed for the end users so that they can download their required dataset in their required format can be accessed by contacting first author.

Fig. 9. Website Hadith database.

Fig. 10 shows an activity diagram about how the user can choose against different options to download their required dataset.

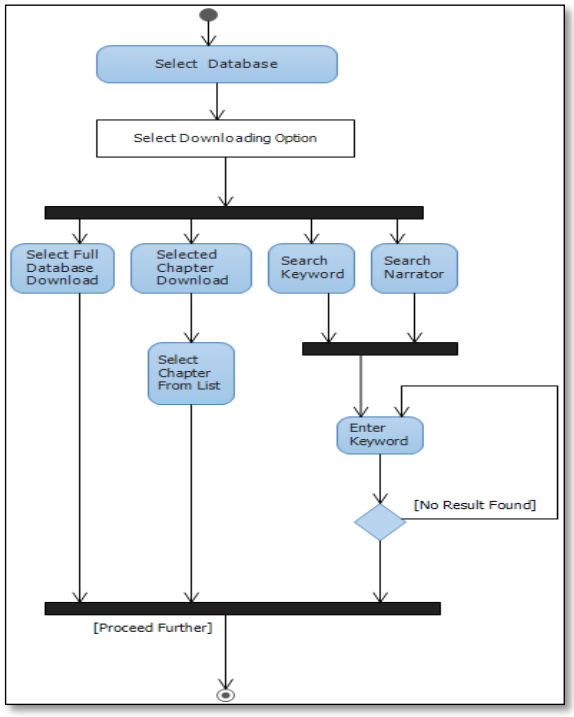


Fig. 10. Activity diagram of downloading the database.

V. Experiments and Results

We use Precision, Recall and F-Measure to measure the performance and accuracy of our methods. Due to the difference in structure of different books, the accuracy level is different. Precision can be given as the ratio of all entities extracted by our system to the correct entities extracted by our system. Its equation can be given as:

$$\text{Precision} = \frac{\text{No of Correct entities Extracted}}{\text{No of all entities Extracted}} \quad (1)$$

Recall can be given as the ratio of total entities extracted by human to the correct entities extracted by our system. Its equation can be given as:

$$\text{Recall} = \frac{\text{No of Correct entities Extracted}}{\text{No of Actual entities extracted}} \quad (2)$$

F1 score can be given as:

$$\text{F1} - \text{Score} = \frac{2 * \text{Recall} * \text{Precision}}{(\text{Recall} + \text{Precision})} \quad (3)$$

On the basis of all the above calculation methods some of the results obtained by our system are given in Table III.

TABLE III. Results of Different Hadith Books

Book Name	Precision	Recall	F1 Measure
Sahih Muslim English	96%	91%	93%
Sahih Bukhari English	99%	99%	99%
Sunan Abudawud	100%	100%	100%
Mawta Imam Malik	100%	100%	100%

In the Sahih Muslim book, our accuracy is less because the structure of Hadiths is complex and noisy and format changes around almost every chapter. Fig. 11 shows the Precision, Recall and F1 Measure rate across different books.

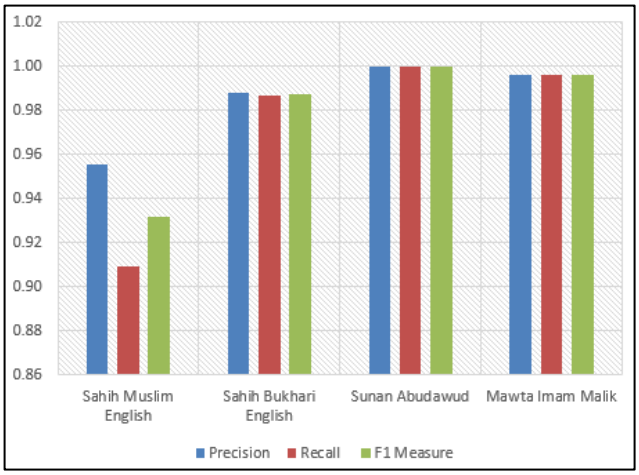
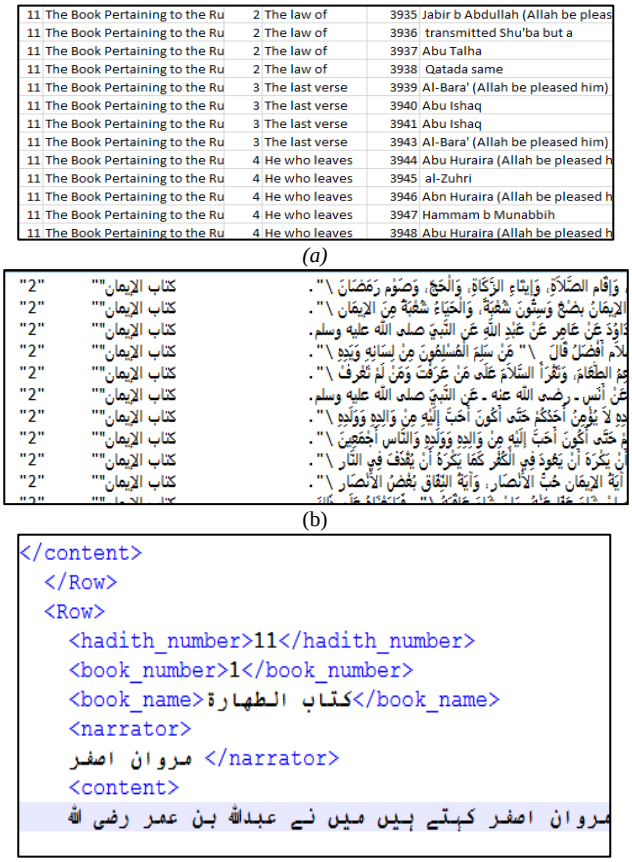


Fig. 11. Precision, recall, F1 measure performance.

Fig. 12(a), (b) and (c) shows Hadith data in different formats. Users can download the dataset in any of these formats. The formats in which user can download dataset are plain Text Format, CSV Format, XLS Format, XML FORMAT.



11	The Book Pertaining to the Ru	2	The law of	3935	Jabir b Abdullah (Allah be pleas
11	The Book Pertaining to the Ru	2	The law of	3936	transmitted Shu'ba but a
11	The Book Pertaining to the Ru	2	The law of	3937	Abu Talha
11	The Book Pertaining to the Ru	2	The law of	3938	Qatada same
11	The Book Pertaining to the Ru	3	The last verse	3939	Al-Bara' (Allah be pleased him)
11	The Book Pertaining to the Ru	3	The last verse	3940	Abu Ishaq
11	The Book Pertaining to the Ru	3	The last verse	3941	Abu Ishaq
11	The Book Pertaining to the Ru	3	The last verse	3943	Al-Bara' (Allah be pleased him)
11	The Book Pertaining to the Ru	4	He who leaves	3944	Abu Huraira (Allah be pleased h
11	The Book Pertaining to the Ru	4	He who leaves	3945	al-Zuhri
11	The Book Pertaining to the Ru	4	He who leaves	3946	Abn Huraira (Allah be pleased h
11	The Book Pertaining to the Ru	4	He who leaves	3947	Hammam b Munabbih
11	The Book Pertaining to the Ru	4	He who leaves	3948	Abu Huraira (Allah be pleased h

(a)

"2"	كتاب الإيمان	وَإِقَامِ الصَّلاَةِ، وَإِيتَاءِ الزَّكَاةِ، وَالْحَجِّ، وَصَوْمِ رَمَضَانَ \".
"2"	كتاب الإيمان	الإِيمَانُ بِضْعٌ وَسِتُّونَ شُعْبَةً، وَالْحَيَاءُ شُعْبَةٌ مِنَ الإِيمَانِ \".
"2"	كتاب الإيمان	دَاوُدَ عَنْ عَامِرٍ عَنْ عَبْدِ اللَّهِ عَنِ النَّبِيِّ صلى الله عليه وسلم.
"2"	كتاب الإيمان	لاَمِ أَفْضَلُ قَالَ \" مَنْ سَلِمَ الْمُسْلِمُونَ مِنْ لِسَانِهِ وَيَدِهِ \".
"2"	كتاب الإيمان	مِ الطَّعَامَ، وَتَقْرَأُ السَّلاَمَ عَلَى مَنْ عَرَفْتَ وَمَنْ لَمْ تَعْرِفْ \".
"2"	كتاب الإيمان	عَنْ أَنَسٍ ـ رضى الله عنه ـ عَنِ النَّبِيِّ صلى الله عليه وسلم.
"2"	كتاب الإيمان	دِهِ لاَ يُؤْمِنُ أَحَدُكُمْ حَتَّى أَكُونَ أَحَبَّ إِلَيْهِ مِنْ وَالِدِهِ وَوَلَدِهِ \".
"2"	كتاب الإيمان	مْ حَتَّى أَكُونَ أَحَبَّ إِلَيْهِ مِنْ وَالِدِهِ وَوَلَدِهِ وَالنَّاسِ أَجْمَعِينَ \".
"2"	كتاب الإيمان	أَنْ يَكْرَهَ أَنْ يَعُودَ فِي الْكُفْرِ كَمَا يَكْرَهُ أَنْ يُقْذَفَ فِي النَّارِ \".
"2"	كتاب الإيمان	آيَةُ الإِيمَانِ حُبُّ الأَنْصَارِ، وَآيَةُ النِّفَاقِ بُغْضُ الأَنْصَارِ \".

(b)

```
</content>
  </Row>
  <Row>
    <hadith_number>11</hadith_number>
    <book_number>1</book_number>
    <book_name>كتاب الطهارة</book_name>
    <narrator>
    مروان اصفر </narrator>
    <content>
    مروان اصفر کہتے ہیں میں نے عبداللہ بن عمر رضی اللہ
```

(c)

Fig. 12. (a) Hadith Database Downloaded in CSV Format, (b) Hadith Database Downloaded in Text Format Arabic Language, (c) Hadith Dataset Downloaded in XML format.

VI. Conclusion

In this paper, we discuss how we prepared hadith repository by applying regular expressions to extract the Hadith data from Multiple Hadith books that are present in different forms, and on both sorts of online & offline data. In the process, we crawled different websites to gather data from the sites directly and have also extracted the data from different sort of files like pdf, doc, etc. we then made a website in WPF to make all the databases downloadable for public. In the future, we plan to analysis data using different data mining and text mining algorithms. In addition, we plan to launch our website to provide online and free access of the dataset repository so that researchers all over the world may download the Hadith data from our website. The output dataset of Hadith can be used in many applications of data mining, text mining and information retrieval. Moreover, there are many fields of NLP which can get benefits from this dataset and can be applied on hadith dataset to extract knowledge in different ways through these datasets.

References

[1] Clinton Cardoza, Rupali Wagh, "Text analysis framework for understanding cyber-crimes," International Journal of Advanced and Applied Sciences, vol. 4, no. 10, pp. 58-63, 2017.

[2] Rehan Khan, Hikmat Ullah Khan, Muhammad Shehzad Faisal, Khalid Iqbal, Muhammad Shahid Iqbal Malik, "An Analysis of Twitter users of Pakistan," International Journal of Computer Science and Information Security, vol. 14, no. 8, 2016.

[3] Altaher, Altyeb, "Hybrid approach for sentiment analysis of Arabic tweets based on deep learning model and features weighting," International Journal of Advanced and Applied Sciences, vol. 4, no. 8, pp. 43-49, 2017.

[4] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, "From data mining to knowledge discovery in databases," AI magazine, vol. 17, no. 3, p. 18, 1996.

[5] Puteri N. E. Nohuddin, Zuraini Zainol , Angela S. H. Lee , A. Imran Nordin , Zaharin Yusoff, "A case study in knowledge acquisition for logistic cargo distribution data Mining framework," International Journal of Advanced and Applied Sciences, vol. 5, no. 1, pp. 8-14, 2018.

[6] Zulfiqar Ali, Waseem Shahzad , Syed Khuram Shahzad, "A review on comparative performance analysis of associative classifiers," International Journal of Advanced and Applied Sciences, vol. 4, no. 6, pp. 96-103, 2017.

[7] Saloot, M. A., Idris, N., Mahmud, R., Ja'afar, S., Thorleuchter, D., & Gani, A, "Hadith data mining and classification: a comparative analysis," Artificial Intelligence Review, vol. 46, no. 1, pp. 113-128, 2016.

[8] Crescenzi, Valter, Giansalvatore Mecca, and Paolo Merialdo, "Roadrunner: Towards automatic data extraction from large web sites.," VLDB, p. 1, 2001.

[9] Michela Becchi ,Anat Bremler-Barr ,David Hay ,Omer Kochba ,Yaron Koral, "Accelerating regular expression matching over compressed HTTP," 2015.

[10] Aldhlan, K.A, Zeki, AM, "Datamining and Islamic knowledge extraction: alhadith as a knowledge resource," 2010.

[11] Fouzi Harrag, Eyas El-Qawasmeh, and Abdul Malik Salman Al-Salman, "Extracting named entities from prophetic narration texts (Hadith)," Berlin, 2011.

[12] Siddiqui, Muazzam Ahmed, Mostafa El-Sayed Saleh, and Abobakr Ahmed Bagais, "Extraction and Visualization of the Chain of Narrators from Hadiths using Named Entity Recognition and Classification," 2014.

[13] F. Harrag, "Text mining approach for knowledge extraction in Sahîh Al-Bukhari," Computers in Human Behavior archive, vol. 30, pp. 558-566, 2014.

[14] Mohanad Jasim Jaber, Saidah Saad, "NER in english translation of hadith documents using classifiers combination," Journal of Theoretical and Applied Information Technology, vol. 84, no. 3, pp. 348-354, 2016.

[15] Ahsan Mahmood, Hikmat Ullah Khan, Zahoor-ur-Rehman, Wahab Khan, "Query based information retrieval and knowledge extraction using Hadith datasets," in 13th International Conference on Emerging Technologies (ICET), Islamabad, 2017.

[16] Aldhaln, Kawther A., Akram M. Zeki, and Ahmed M. Zeki, "Knowledge extraction in Hadith using data mining technique," 2012.

[17] Fouzi Harrag, Aboubekeur Hamdi-Cherif, and Eyas El-Qawasmeh, "Vector space model for Arabic information retrieval—application to "Hadith" indexing," 2008.

[18] Nursyahidah Alias, Zulhilmi Mohamed Nor, Nurazzah Abdul Rahman, "Searching Algorithm of Authentic Chain of Narrators' in Shahih Bukhari Book," MALAYSIA, 2016.

[19] Shatnawi MQ, Abuein QQ, Darwish O, "Verification Hadith Correctness in Islamic Web Pages Using Information Retrieval Techniques," 2011.

[20] K. Jbara, ". "Knowledge discovery in Al-Hadith using text classification algorithm," Journal of American Science, vol. 6, no. 11, pp. 409-19, 2010.

[21] Azmi AM,Alkhalifah F, Alsaeed A, Barnawi Y, "Using non-conventional search schemes to retrieve Hadiths," 2014.

[22] Maheen Akhter Ayesha, Sahar Noor, Muhammad Ramzan, Hikmat Ullah Khan, Muhammad Shoaib, "Evaluating Urdu to Arabic Machine Translation Tools," International Journal of Advanced Computer Science and Applications, vol. 8, no. 10, pp. 90-96, 2017.

[23] Boella M, Romani FR, Al-Raies A, Solimando C, Lancioni G, "The SALAH Project: Segmentation and Linguistic Analysis of Hadith Arabic Texts," 2011.

[24] Khan HU, Saqlain SM, Shoaib M, Sher M, "Ontology Based Semantic Search in Holy Quran," International Journal of Future Computer and Communication, vol. 2, no. 6, 2013.

[25] Muhammad Shoaib M, Yasin MN, Khan HU, Saeed MI, Khiyal MS, "Relational WordNet Model for Semantic Search in Holy Quran," 2009.

[26] Kais Dukes, Tim Buckwalter , "A Dependency Treebank of the Quran using Traditional Arabic Grammar," 2010.

[27] M. Y. Alfi, "An Applied Linguistics Approach to Improving the Memorization of the Holy Quran: Suggestions for Designing Practice Activities for Learning and Teaching.," Journal King Saud Univ, vol. 16, pp. 1-32, 2004.

[28] Mohammad Alqahtani, Eric Atwell, "Arabic Quranic Search Tool Based on Ontology," Natural Language Processing and Information Systems, vol. 9612, pp. 478-485, 2016.

[29] Ali, Imran, "Application of a Mining Algorithm to Finding Frequent Patterns in a Text Corpus: A Case Study of the Arabic," International Journal of Software Engineering and Its Applications, vol. 6, no. 3, 2012.

[30] Chee Yong Chan, Minos Garofalakis, Rajeev Rastogi, "Indexed Regular Expression Matching," Springer US, pp. 1-6, 2014.

[31] Brodie, Benjamin C., David E. Taylor, and Ron K. Cytron, "A scalable architecture for high-throughput regular-expression pattern matching," ACM SIGARCH Computer Architecture News, vol. 34, no. 2, 2006.

[32] Garrett, Jesse James, "Ajax: A new approach to web applications," 2005.

[33] Fang Yu, Zhifeng Chen, Yanlei Diao, T.V. Lakhsman, Randy H.Katz, "Fast and memory-efficient regular expression matching for deep packet inspection," 2006.