Hadith data mining and classification: a comparative analysis

Mohammad Arshi Saloot, Norisma Idris, Rohana Mahmud, Salinah Ja'afar, Dirk Thorleuchter & Abdullah Gani

Artificial Intelligence Review
An International Science and
Engineering Journal

ISSN 0269-2821
Volume 46
Number 1

Artif Intell Rev (2016) 46:113-128
DOI 10.1007/s10462-016-9458-x

Springer

Artif Intell Rev (2016) 46:113–128
DOI 10.1007/s10462-016-9458-x

Hadith data mining and classification: a comparative analysis

Mohammad Arshi Saloot[1] · **Norisma Idris**[1] · **Rohana Mahmud**[1] · **Salinah Ja'afar**[1] · **Dirk Thorleuchter**[2] · **Abdullah Gani**[1]

Published online: 8 January 2016

Abstract Hadiths are important textual sources of law, tradition, and teaching in the Islamic world. Analyzing the unique linguistic features of Hadiths (e.g. ancient Arabic language and story-like text) results to compile and utilize specific natural language processing methods. In the literature, no study is solely focused on Hadith from artificial intelligence perspective, while many new developments have been overlooked and need to be highlighted. Therefore, this review analyze all academic journal and conference publications that using two main methods of artificial intelligence for Hadith text: Hadith classification and mining. All Hadith relevant methods and algorithms from the literature are discussed and analyzed in terms of functionality, simplicity, F-score and accuracy. Using various different Hadith datasets makes a direct comparison between the evaluation results impossible. Therefore, we have re-implemented and evaluated the methods using a single dataset (i.e. 3150 Hadiths from Sahih Al-Bukhari book). The result of evaluation on the classification method reveals that neural networks classify the Hadith with 94 % accuracy. This is because neural networks are capable of handling complex (high dimensional) input data. The Hadith mining method that combines vector space model, Cosine similarity, and enriched queries obtains the best accuracy result

✉ Mohammad Arshi Saloot
phd_siamak@yahoo.com

Norisma Idris
norisma@um.edu.my

Rohana Mahmud
rohanamahmud@um.edu.my

Salinah Ja'afar
b1salina@um.edu.my

Dirk Thorleuchter
dirk.thorleuchter@int.fraunhofer.de

Abdullah Gani
abdullah@um.edu.my

[1] University of Malaya, 50603 Kuala Lumpur, Malaysia

[2] Institute of Fraunhofer INT, Appelsgarten 2, 53879 Euskirchen, Germany

(i.e. 88 %) among other re-evaluated Hadith mining methods. The most important aspect in Hadith mining methods is query expansion since the query must be fitted to the Hadith lingo. The lack of knowledge based methods is evident in Hadith classification and mining approaches and this absence can be covered in future works using knowledge graphs.

Keywords Review · Comparison · Islamic knowledge · Hadith · Classification · Data mining

1 Introduction

Many researchers around the world, such as the Natural Language Processing research group, part of the Institute for Artificial Intelligence and Biological Systems (I-AIBS) at Leeds University, have undertaken significant experiments in respect of Quranic Arabic in the past decade. They have enriched the Quran text with annotations, captured its linguistic structure, and provided tools to perform automated, objective inference and querying (Atwell et al. 2011). One of the advantage of applying natural language processing (NLP) tasks on the Islamic text is the implementation of intelligent systems which can answer any question with data from the Quran and Hadith, and can assist Muslim and non-Muslim societies, to apprehend the Quran and Hadith. However, another important Islamic source, Hadith, has been overlooked by most academics in computer science. This paper offers a review of the classification and data mining tasks concerning Hadith texts.

Hadith is derived from the Arabic word "Hadatha" meaning news and story. There are different branches and various types of schools in Islam such as Sunni, Shia, and Sufi. Different branches of Islam refer to different collections of hadith. According to the Sunni branch of Islam, a Hadith is any speech, discussion, action, approval, and physical or moral description attributed to the Prophet Muhammad, whether supposedly or truly (Batyrzhan et al. 2014). On the other hand, Shias add in their Hadith collections not only statements associated to the Prophet Mohammad but also statements associated to their Imams, whom they regard as infallible leader (Fattāhizādeh and Afshāri 2010). In order to take a close look at Hadith, we have to know its components as displayed in Fig. 1: Matn (the central text), Isnad (chain of narrators), and Taraf (the beginning phrase(s) of the Hadith), which indicate the sayings, actions or characteristics of the Prophet. The authenticity of the hadith depends on the reliability of its components, and the linkage among them.

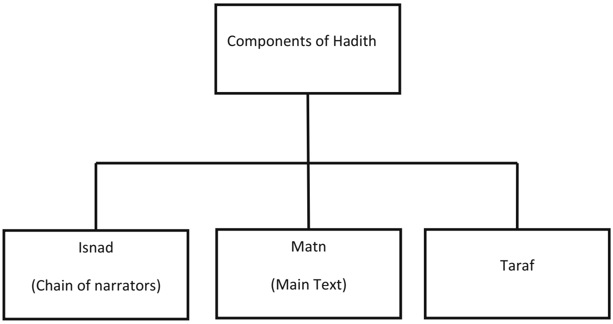

Fig. 1 Structure of Hadith

Considering that Prophet Muhammad passed away 14 centuries ago, how can we be sure that the sayings and doings we attribute to him are unaffected and uninterrupted? To someone unaccustomed to the science of Hadith, the compilation of Hadith may seem inaccurate or deceptive. There is a systematic approach in Islamic science that helps us to distinguish a valid Hadith from an invalid one, named "Hadith science" (Melchert 2002). As time passed, more narrators were associated with each Isnad, and so the situation required solid rules and criteria for the acceptance of Hadith; these rules and criteria are known as Mustalah Al-hadith (the classification of Hadith) (Lacroix 2008). The eventual verdict on a Hadith would determine its category: (1) Sahih refers to an authentic Hadith, (2) Hasan (meaning "good") describes a Hadith the authenticity of which is not confirmed as Sahih Hadith, but is acceptable, (3) Daif refers to a weak Hadith, and (4) Maudu is a term applied to fabricated or forged Hadiths. Surprisingly, the principals of Hadith science have merged to other areas of science, such as digital evidence authentication. Yusoff et al. (2010) used primary elements of Hadith science in digital evidence authentication in the following areas: content verification, transmitter reliability, transmission reliability, and change of custody.

Since the Quran and Hadith serve as the two fundamental scriptural sources of Islamic beliefs, there are many studies on them from a computer science perspective. In this paper, we classify the published works about Hadith into two major groups. (1) Studies on classification: the most published Hadith studies focus on the classification. In general, automatic classification is used in many applications, such as spam filtering, email routing, language identification, genre classification, sentiment analysis, and article triage (Sebastiani 2002). Muslim scholars started practicing classification of Islamic texts in ancient times in order to group Hadiths according to their topics. Automatic Hadith classification is an innovative research topic that uses different methods, such as decision trees, support vector machine (SVM), artificial neural networks, K-nearest neighbor (KNN), and Naive Bayes probabilistic classifier (Yang and Liu 1999). (2) Studies on data mining: Data mining plays an important role in knowledge discovery (Cabena et al. 1998), and, therefore, it is used to explore Hadith and to bring forth the desired data.

As the religious language of Islam, Arabic is the universal language of 1.5 billion Muslims around the world. Therefore, most of the Islamic and Hadith resources are in Arabic. The Arabic language is one of the primeval languages. It is still extensively used today and is the native language of millions of citizens in the Arab countries. It is a Semitic middle eastern language and is written from right to left (Al-tarawneh et al. 2014). The archaic Arabic writing system was consonantal. Each letter in the Arabic alphabet has represented a single consonant since ancient times. However, late in the seventh century, the Arabic diacritics, which are graphical symbols that discriminate between the variety of pronunciations of the consonants, were invented by "Abu Al-Aswad Al-Du'ali". Nevertheless, they are very often eliminated from the written text of today. Arab readers could discern words with the same writing form via its context (Abuzeina 2011).

Dividing the input text into desired fractions is usually the initial stage in most text processing tasks. These could be sentences, numbers, words, characters, or any other useful fraction. Each fraction is called a token and the process is called tokenization. In Arabic, a token may specify a whole grammatical phrase for example, "وسنساعدهم" (meaning: "and we shall help them") (Attia et al. 2010). One of the most effective elements in distinguishing sentences or token boundaries is punctuation marks. They emerged in the Arabic writing system in 1912. In fact, punctuation usage is not persistent in the Arabic language.

This article is a review of all the studies on Hadith in classification and data mining spheres that have been published in prestigious journals or conference proceedings. Since Hadith and Islamic subjects are topics of interest to Arabic language scholars, there are many relevant articles in the Arabic language, such as Alrazou (2004, 2008). In order to focus on research that is understandable to all readers, only English-language articles were included. The remainder of this paper is organized as follows: Sect. 2 presents related works on Hadith classification. Section 3 discusses the data mining in Hadith. Then, our experiment on a comparison of different methods is illustrated in Sect. 4. Finally, Sect. 5 concludes this paper with a brief summary and future works.

2 Hadith classification

Kabi et al. (2005) proved that Hadith classification could be handled via a simple method with reasonable accuracy. They classified Arabic Hadiths into eight chapters of Sahih Al-Bukhari by calculating the term frequency. After removing Isnad and stop words, each word was converted to its root form by a stemmer system. In Sahih al-Bukhari, the same Hadith might exist in more than one chapter. In that case, their system displayed two subjects with the highest ranks. In order to perform term weighting, the TF-IDF method was employed and trained by 120 Hadiths. Testing the system with 80 Hadiths presented 83.2 % accuracy.

In a research proposal, Kamsin et al. (2014) emphasized the importance of an automatic authentication system for the Quran and Hadith in order to combat fake versions of the Quran and Hadith in the virtual sphere. Ghazizadeh et al. (2008) built a fuzzy expert system based on a set of rules and experts' views. In order to make inference, two inference engines were designed. The first engine produced the rank of each narrator and passes it to the second inference engine. The product of the second engine is the Hadith validation rate. Their system was tested using the KAFI dataset to rank Hadiths with unknown, weak, goodness, reliable, and right regarding their authenticity rate. The system achieved 94 % accuracy via a combination of triangular fuzzifier, singleton fuzzifier, multiplication inference, Mamdani multiplication inference, and average defuzzifier.

Artificial neural networks (ANNs) are one of the methods to perform classification. Harrag and El-Qawasmeh (2009) employed an ANN method to classify Hadiths. In their classification approach—apart from text pre-processing—using singular value decomposition (SVD) is the first step, which is an effective data cleaning process. There are 739 unique vocabulary in the dataset. Each feature has referred to one vocabulary. The SVD converts the sparse and high dimensional vector to a 200 dimensional vector of document-terms weights. They used the Prophetic encyclopedia (The Encyclopedia of the Nine Books for the Honorable Prophetic Traditions! 1997), which includes 453 documents distributed over 14 categories: faith, Quran, knowledge, crimes, Al-Jihad, good manners, past generations, biography, judgments, worships, behaviors, food, clothes, and personal states. A three layer feed-forward neural network with hyperbolic tangent (tanh) activation function in the hidden layer, followed by a linear output layer was trained using a back-propagation algorithm. The obtained recall, precision, and F1-score for the Hadith category prediction was 87, 90, and 88 %, respectively.

In a decision tree classifier, leaves refer to class labels and branches refer to the coincidence of features that point to particular labels (Maazouzi and Bahi 2012). In addition, a top-down algorithm is required to traverse the tree and predict classes. Harrag et al. (2009) illustrated a classification experiment using 453 Hadiths distributed over 14 groups from the Prophetic encyclopedia via ID3 classifier (Flachsbart et al. 1994). The preprocessing phase consisted of converting the document to plain text, removing stop words, and stemming. After preprocessing, a vector was constructed consisting of all the terms in the Hadith texts.

Then, a vector dimension was reduced to 1938 based on some specific criteria and a weight was calculated for each dimension using term frequency (TF). The evaluation in the testing phase yielded 38 % recall, 47 % precision, and 40 % F-score. There were a variety of misclassification laps because of the nature and characteristics in the Hadith documents.

Sahih Al-Bukhari is partitioned into many chapters regarding the subject of Hadiths, but Alkhatib (2010) only considered eight of them in the experiment: (1) knowledge, (2) faith, (3) hajj, (4) praying, (5) obligatory charity, (6) eclipse prayer, (7) fasting, and (8) good manners. Alkhatib (2010) examined four classification algorithms: (1) Naive Bayes (Rennie et al. 2003), (2) Rachio algorithm (Ragas and Koster 1998), (3) K-nearest neighbor (KNN) (Zhang and Zhou 2005), and (4) support vector machine (SVM) (Hsu and Lin 2002). To calculate the relative frequency for each word in the document, the term frequency-inverse document frequency (TF-IDF) method was employed (Aizawa 2003). For training purposes, 1,350 Hadiths were used and 150 Hadiths for testing the accuracy of the classification methods. The average recall of all the methods was 100 % but the precision of Rachio was 67.11 %, Naive Bayes 66.55 %, KNN 66.55 %, and SVM was 63.36 %. Therefore, the Rachio algorithm classified Hadiths with the highest precision level.

Harrag et al. (2011b) provided an evaluation study of several stemming methods for Hadith text categorization. The study examined the dictionary-lookup stemming, the root-based stemming, and the light stemming as a feature reduction stage. The dictionary-lookup stemming performed the truncation via two sets of resources: a basis of affixes and a dictionary of words with their corresponding roots. The root-based stemmers execute pattern matching to discover the root of the words. The light stemming strips off some suffixes and/or prefixes, without recognizing patterns. The ANN and the SVM are chosen for the categorization stage. The experiment was performed on 453 Hadith texts from the Prophetic encyclopedia, which was distributed over 14 categories. After performing preprocessing routines and stemming, the data were represented by four vectors: un-stemmed vector with 4,055 dimensions, light stemmed vector with 2536 dimensions, root-based stemmed vector with 1063 dimensions, and dictionary-lookup stemmed vector with 739 dimensions. The experimental results proved that the ANN method was better than SVM in terms of the F1-score. The obtained F1-score without stemming for ANN was 42 % and for SVM it was 44 %, although it was enhanced after using the stemmers. The dictionary-lookup stemming achieved the highest accuracy (i.e. an F1-score of 50 %) compared to the root-based stemming and light-stemming methods for the ANN classifier. The light stemming achieved the highest accuracy (i.e. an F1-score of 48 %) compared to root-based stemming and dictionary-lookup stemming methods for the SVM classifier. In addition, the shortened size of the vectors caused less computational efforts for both classifiers.

Aldhaln et al. (2012a, b) classify Hadith into four major classes Sahih, Hasan, Daif, and Maudo. In order to form training and testing datasets, 999 Hadiths were collected from three books: Sahih Al-Bukhari, Jamiu Al-Termithi, and Silsilat Al-Ahadith Al-Daeifah Wal Al-Mawdhuah. Since the dataset was composed from different books, data pre-processing was performed to diminish redundancy and to make the style of Hadith Isnad homogeneous. The features of the Hadiths were determined according to the five principles in Hadith science: (1) distinguished narrators for their candor, (2) distinguished narrators for their veracity, (3) no interrupting in the Isnad, (4) no abnormal phrase in the Hadith Matn, and (5) no faultiness in the Hadith Matn. They presented a novel method to handle the missing data in the Hadith dataset. The classification task was accomplished via two different methods: decision trees and Naive Bayes. However, the Naive Bayes classifier produced better results with 97.6 % recall and 97.597 % accuracy. They also published an extended version of the paper that includes more technical details about the same experiment in a journal article

(Aldhaln et al. 2012). In a comparative study, Al-Kabi and Al-Sinjilawi (2007) proved that the Naive Bayes classifier produced the best results compared to other classifiers in terms of Sahih Al-Bukhari.

Najeeb (2014) proposed a new classification approach that distinguishes between authenticated (Sahih) and weak (Daif) Hadiths. Associative classification (Hu et al. 1999) was used that assimilated classification and association rule mining (ARM), also called CBA (classification based on associations). The ARM aimed to discover the relationship between the features in order to define a set of classification rules. Although, the functionality of the CBA on the Hadith domain has been confirmed, no explicit accuracy rate has been reported.

Table 1 gives an overview of the above mentioned classification approaches and depicts them together with the used methods, categories, and data as well as the corresponding evaluation results.

3 Data mining

Karim and Hazmi (2005) conducted a qualitative data analysis through interviewing Malaysian postgraduate students in order to evaluate information about Hadiths on the Internet. The result showed that almost all of the participants consider the Internet as a convenient Hadith resource, however; there is a risk of using faulty Hadiths. Shatnawi et al. (2012) explained an experiment that contained two major steps: (1) retrieve Hadiths from web pages and (2) verify correctness of the retrieved Hadiths. They used a database provided by Sheikh Al-Albani that contained more than 17,000 Hadith texts along with their authentication degrees. Shatnawi et al. (2012) illustrated how to tokenize the database and remove stop words. Except for the 28 Arabic alphabets, all the characters were removed including Arabic vowel marks. Finally, a positional index was built, which contained more than 56,000 terms. In order to extract the Hadith texts from web pages, a Java HTML cleaner eliminated the HTML codes from the web pages. Then, four contiguous words from the web page were compared with the Hadith positional index to detect the Hadith text. When all the Hadith texts were fetched, each one was looked up in the database to determine their degree of authenticity. When five web pages containing Hadith texts were randomly chosen (encompassing 63 Hadith texts), 76.1 % precision and 42.1 % recall were achieved.

AthenTique is a text mining tool to search a query from a Hadith dataset (Harrag et al. 2008; Harrag and Hamdi-Cherif 2007). It displays a list of relevant Hadiths sorted by the extent of similarity. AthenTique works based on the vector space model (VSM) (Salton et al. 1975). The VSM refers to a concept that stores all information as indexes before measuring the similarity of the query and main text. The first step is the Hadith morphological stemming based on a root dictionary. After the preprocessing is performed on all Hadiths, the term weighting process and indexing can begin via the TF-IDF method. In addition, each query is processed in the same way as the Hadith dataset. Then, the similarity is measured between the query and Hadiths with the help of the cosine measure technique. The Hadith retrieval is performed in two rounds: (1) all terms from the top five relevant documents are ordered by their weight of their relevancy to the original query. (2) the top ten terms from the retrieved documents are integrated to the original query to form an enriched query, while new terms have lower weights than initial query terms. A conducted experiment with 60 Hadiths obtained 66 % precision and 80 % recall.

A conventional information retrieval system fetches one or more documents based on a query. Each document is usually so lengthy that users cannot traverse the whole retrieved document. Topic segmentation can be utilized in information retrieval, in which the task is to

Table 1 Overview on classification approaches

No.	Approach published by	Methods	Categories	Data source	Performance
1	Kabi et al. (2005)	TF-IDF	Eight subjects	200 Hadiths from Sahih Al-Bukhari	Accuracy: 83 %
2	Ghazizadeh et al. (2008)	Fuzzy expert system	Unknown, weak, goodness, reliable, and right	KAFI Dataset	Accuracy: 94 %
3	Harrag and El-Qawasmeh (2009)	Term indexing, artificial neural network, singular value decomposition	14 subjects such as: faith, Quran, knowledge, crimes, good manners and judgments	Prophetic encyclopedia	Recall: 87 %, Precision: 90 %, F1-score: 88 %
4	Harrag et al. (2009)	Term frequency, ID3 decision tree classifier,	14 subjects such as: faith, Quran, knowledge, crimes, good manners and judgments	Prophetic encyclopedia	Recall: 38 %, Precision: 47 %, F1-score: 40 %
5	Alkhatib (2010)	TF-IDF, Rachio algorithm	Eight subjects	1350 Hadiths from Sahih Al-Bukhari	Precision: 67 %
6	Harrag et al. (2011b)	TF-IDF, dictionary lookup stemming, and artificial neural network	14 groups such as: faith, Quran, knowledge, crimes, good manners and judgments	Prophetic encyclopedia	F1-score: 50 %
7	Aldhaln et al. (2012a, b) and Aldhaln et al. (2012)	Decision trees and Naive Bayes	Sahih, Hasan, Daif and Maudo	999 Hadiths from three books: Sahih Al-Bukhari, Jamiu Al-Termithi, and Silsilat Al-Ahadith Al-Daeifah Wal Al-Mawdhuah	Recall: 97 %, Accuracy: 97 %
8	Najeeb (2014)	Term indexing, associative classification	Sahih and Daif	Not reported	Not reported

split a document into a set of topically meaningful segments. Harrag et al. (2009) described two different experiments upon a unique dataset: information retrieval with segmentation and without segmentation. For information retrieval without segmentation, roots were extracted from the query text to calculate their weight by TF-IDF. In fact, TF-IDF gave a biased weight to the rare terms in the Hadith dataset. Then, indexing the terms of relevant Hadiths and the count of the relevance weight were determined by their specific equations. Finally, the system displayed two sets of Hadiths: relevant and irrelevant. In the experimentation phase, the average recall was 54 %, and the precision was 41 % using a dataset that contained 453 Hadiths. For information retrieval with segmentation, the C99 segmentation algorithm (Choi 2000) was utilized. For evaluation, a segmented text was prepared by paired agreements among seven human experts using Kappa coefficient. Finally, the system displayed a list of segments sorted according to their relevancy. After using topic segmentation, there was an improvement of 14 % for recall, and 51 % for precision.

As emphasized by Aldhaln et al. (2010) in their review article, various types of information can be derived as a knowledge form of Hadith, including Islamic legislative, Islamic military, and Hadith classification. Jbara (2010) proposed a text mining system to retrieve a Hadith category in response to a query. In this case, 1321 Hadiths from the Sahih Al-Bukhari book were prepared to train and test the system distributed over thirteen groups: faith, knowledge, praying, call to prayer, eclipses, alms, good manners, fasting, medicine, food, pilgrimage, grievance, and virtues of the Prophet Mohammad. The experiment included three phases: The first phase applied preprocessing, which consisted of removing Isnad, tokenization, removing punctuation and diacritical marks, removing stop words, and stemming. The second phase was used for training, in which the feature matrix (vocabularies are features) was constructed using TF-IDF. The third phase enabled classification in which the resulting training dataset of the previous phase was used. In addition, query features weight the computation and query expansion performed in the third phase. Finally, the category prediction was accomplished with a similarity coefficient table and the maximum cumulative similarity method, in which 45 % precision and 49 % F1-score were obtained.

Bilal and Mohsin (2012) introduced a cloud based and distributed rule-based expert system that uses Hadith science to classify them as authentic and unauthentic, known as Muhadith. Queries can be provided via a web interface as implemented in Muhadith. The five main modules of the Muhadith are: (1) inference engine: a set of if-then-else statements that constitutes the rules, (2) knowledgebase: binary decision tables for representing knowledge, (3) parser and fact extractor: used to pars the queries provided by the users in a first step and used to extract relevant information concerning the provided query, (4) explanation facility: used to provide details to users about how and why a conclusion has been drawn, and (5) database. Muhadith is developed as a service-oriented architecture (SOA) based Cloud expert system accessible through the web. No experimental result or evaluation method is reported in the paper.

The named-entity extraction technique was employed to identify useful entities from a Hadith dataset (Harrag et al. 2011a; Harrag 2014). As mentioned earlier, Sahih Al-bukhari was divided into more than 91 chapters according to the subject of the Hadiths. Each chapter was subdivided into several sections, with a total of 3882 sections. Sahih Al-bukhari contains more than 9000 Hadiths, and each Hadith contains a Hadith number, an Isnad (chain of narrators), a main text (Matn), and beginning phrase (Taraf). To transform the unstructured text into a semi-structured text file, they defined a process to detect desired entities: chapter number, chapter title, section number, section title, Hadith number, Isnad, Matn, Taraf, and date. They implemented a model using a finite state transducer (Roche and Shabes 1997) in the form of automata. The automata were represented by a set of states and transitions between these states, while a text was linked to each state. The automata turned a sequence of vectors (i.e. words) into a sequence of symbols (i.e. entities). The model achieved encouraging precision (71 %), recall (39 %), and F1-score (52 %) rates. Although it performed well in identifying number entities like chapter number, section number, and Hadith number, it performed poorly in detecting the date entity because the dates were written in alphabetical format.

The halal term refers to any action or object that is permitted or allowed according to Islamic Law. Many scripts about halal products are available through resources, such as web pages, e-books, and magazines. The end-user may inquire about halal-related data via query phrases to fetch a list of relevant documents. To address this scenario, Hanum et al. (2014) scrutinized topic analysis techniques, that is, latent semantic indexing (LSI) (Papadimitriou et al. 1998) and frequency-based inverted indexing. The experiment was performed upon four Malay translated Hadith documents containing 436 words, and 36 other Malay language documents. In the dataset, 16 documents encompassed various aspects of halal-related subjects. To obtain a vector of words, after tokenization and eliminating stop words, all tokens were converted to their root form using a Malay language stemmer. The similarity between the query and the documents was measured using the cosine similarity technique (Tata and Patel 2007). Five sets of queries, which contained words about halal products, were constructed. In order to evaluate the success of the technique, the dataset was manually analyzed and a list of relevant judgments was compiled. The experiment proves that utilizing LSI gives better results than frequency analysis, that is, 37 % precision and 100 % recall at best. However, LSI needs high computational resources to deal with large documents, such as web pages.

Table 2 gives an overview on the above-mentioned data mining approaches and depicts them together with the used methods, categories, and data as well as the corresponding evaluation results.

4 Comparison

The discussed approaches in Sects. 2 and 3 used different data sources in their experiments. Very often, Sahih Al-Bukhari is used, however, evaluation is based on different Hadiths within the book and various versions and editions of Sahih Al-Bukhari are used. This makes a direct comparison between the evaluation results of the different approaches impossible. In addition, they used different evaluation measures. Therefore, we implemented and evaluated the classification and data mining approaches using the same dataset—an Arabic version of Sahih Al-Bukhari consisting of 3150 Hadiths distributed over 14 groups. The dataset divided to two sections: (1) training dataset consisting of 2520 Hadiths, (2) testing dataset consisting of 630 Hadiths. Both training and testing datasets include all 14 groups. We implemented the mentioned classification and data mining approaches and evaluated them using the accuracy measure.

In order to have a precise comparison, the same preprocessing procedure was followed for all experiments. The first step in the preprocessing phase considered the document preparation in which Isnad and other unrelated data were eliminated. Unlike English text, all Arabic natural language processing tasks need a normalization process. We used AraNLP to eliminate diacritics and punctuation and to address alphabet inconsistent variations (Althobaiti et al. 2014). Tokenization is a mandatory and essential step in natural language processing. After the normalization, we employed a simple white space Tokenizer. Then, stop words were removed through searching in a list of Arabic stop words, which consisted of more than 6000 terms. In stemming, a rule-based Arabic stemmer was utilized, which was proposed by

Table 2 Overview of data mining approaches

No.	Approach published by	Methods	Extracted data	Data source	Performance
1	Shatnawi et al. (2012)	Web data extraction, and positional index	Extract Hadiths from web and determine their authenticity degrees: Sahih, Hasan, Daif, and Maudo	17,000 Hadiths from Sahih Al-Bukhari	Precision: 76 %, Recall: 42 %
2	Harrag et al. (2008) and Harrag and Hamdi-Cherif (2007)	Vector space model, TF-IDF, cosine similarity, enriched query	Relevant Hadith	60 Hadiths	Precision: 68 %, Recall: 80 %
3	Harrag et al. (2009)	C99 segmentation, TF-IDF	Relevant and irrelevant Hadiths	453 Hadiths	Precision: 92 %, Recall: 68 %
4	Jbara (2010)	TF-IDF and similarity coefficient table	13 Hadith groups e.g. faith, knowledge, praying, eclipses, alms, good manners, fasting	1321 Hadiths from the Sahih Al-Bukhari	Precision: 45 %, F1-score: 49 %
5	Bilal and Mohsin (2012)	Cloud based and distributed rule-based expert system	Authentic versus unauthentic	Not reported	Not reported
6	Harrag et al. (2011a) and Harrag (2014)	Finite state transducer	Chapter number, chapter title, section number, section title, Hadith number, Isnad, Matn, Taraf, and date	2602 Hadiths from Sahih Al-Bukhari	Precision: 71 %, Recall: 39 %, and F1-score: 52 %
7	Hanum et al. (2014)	LSI, TF-IDF, and cosine similarity	Halal related Hadiths	16 Malay translated Hadith documents	Precision: 37 %, Recall: 100 %

Al Kharashi and Al Sughaiyer (2002). Furthermore, all the preprocessing steps were repeated for the queries in the data mining tasks.

According to all the mentioned approaches, after the preprocessing stage, term vectors should be compiled. Some approaches follow a simple term indexing procedure; however, TF-IDF has more followers. Therefore, we built vectors according to Eq. 1:

$$TFIDF(w,d) = TF_{w,d} \cdot IDF_{w,d} = TF_{w,d} \cdot \left[\left[\log_2 \frac{ND}{DF_w} \right] + 1 \right] \tag{1}$$

where w is a word, d is a Hadith document, $TF_{w,d}$ is the number of occurrences of w in d, and ND is the total number of Hadith documents in the dataset, and DF_w is the number of Hadith documents that contain w.

4.1 Classification

We could not implement and eavaluate the approach proposed by Aldhaln et al. (2012) and Ghazizadeh et al. (2008) because they followed the Hadith science structures, which focused on the chain of narrators and experts' view instead of terms frequency. The classification approaches were trained and tested on 14 groups of Hadith: personal matters, ethics, food, forbidden matters, previous nations, faith, felony, jihad, behavior, worship, science, Quran, ornamental dress, and transaction.

The results obtained in our experiment are depicted in Table 3, which is sorted according to their rank of accuracy. The obtained accuracy is different from what is reported in the original articles due to using a different Hadith dataset. The approach presented by Harrag et al. (2011b), which used the dictionary lookup stemming for feature reduction, and ANN for classification, achieved the highest accuracy (94 %). The second rank belonged to the approach introduced by Harrag and El-Qawasmeh (2009), which attained 90 % accuracy. The approaches that were placed in the first and the second ranks used ANN to accomplish their classification. However, Harrag and El-Qawasmeh (2009) utilized the SVD technique for the feature reduction.

The approach presented by Alkhatib (2010) achieved 83 % accuracy, which employed the Rachio algorithm for the classification. The approach introduced by Harrag et al. (2009), which utilized the ID3 decision tree classifier without any feature reduction technique, obtained 81 % accuracy. Najeeb (2014) suggested using associative classification, which only achieved 72 % accuracy. Finally, Kabi et al. (2005) were placed in the lowest rank with 70 % accuracy because they only accomplished term indexing via TF-IDF and simply compared the vectors and simply group similar vectors.

4.2 Data mining

The reported data mining experiments in Sect. 3 were re-implemented and re-evaluated except for four: (1) the approach proposed by Harrag (2014) was designed to work upon the specific structure of their dataset, (2) the approach presented by Harrag et al. (2009) required segmented documents along with human judgment, (3) the system introduced by Bilal and Mohsin (2012) classified hadiths based on a set of rules, while the complete list of rules were not reported in the paper, and (4) the approach presented by Shatnawi et al. (2012) focused on parsing the web pages in order to use the web as a dataset.

The results obtained in our experiment are depicted in Table 4, which is sorted according to their rank of accuracy. In order to perform evaluation, human experts manually compiled a ranked list of relevant Hadiths regarding 110 specific queries. Four example queries are shown

Table 3 The obtained results from the classification experiments

Rank No.	Approach published by	Methods	Performance
1	Harrag et al. (2011b)	Dictionary-lookup stemming, and artificial neural network	Accuracy: 94 %
2	Harrag and El-Qawasmeh (2009)	Artificial neural network, and singular value decomposition	Accuracy: 90 %
3	Alkhatib (2010)	Rachio algorithm classifier	Accuracy: 83 %
4	Harrag et al. (2009)	ID3 decision tree classifier	Accuracy: 81 %
5	Najeeb (2014)	Associative classification	Accuracy: 72 %
6	Kabi et al. (2005)	TF-IDF	Accuracy: 70 %

Table 4 The obtained results from the data mining experiments

Rank No.	Approach published by	Methods	Performance
1	Harrag et al. (2008) and Harrag and Hamdi-Cherif (2007)	Vector space model, cosine similarity, and enriched query	Accuracy: 88 %
2	Hanum et al. (2014)	LSI, cosine similarity	Accuracy: 86 %
3	(Jbara 2010)	Similarity coefficient table	Accuracy: 73 %

Table 5 Pre-defined queries

No.	English translation of query	Number of relevant Hadiths
1	What are the non-halal foods	41
2	Halal	112
3	Explain the history of worship	24
4	How to judge mothers	9

in Table 5. The outputs of the re-evaluated approaches—relevant hadiths—were compared with the experts' Hadiths to calculate the accuracy. The obtained accuracy was different from that reported in the original articles due to the use of a different Hadith dataset.

The approach proposed by Harrag et al. (2008) was placed in the first rank, and it was proven that by using VSM and cosine similarity, and making an enriched query, we could achieve 88 % accuracy. Although the approach proposed by Hanum et al. (2014) was designed to classify Malay translated Hadith, it showed significant results for Arabic Hadiths. By employing the LSI and cosine techniques, we achieved compelling accuracy (86 %). The approach presented by Jbara (2010) came in last place. It only used a simple similarity coefficient table to compare the query with Hadiths, which led to only 73 % accuracy.

5 Discussion and conclusion

The Quran and Hadith are two important sources in the world of Islam. Since Hadiths have unique linguistic features, such as ancient Arabic language and story-like text, there is a demand for specific natural language processing methods. This review analyze all studies on Hadith in classification and data mining fields. There are two major aspects in the reported approaches: (1) Features in the experiments are based on aspects of words. For example, the approach of Alkhatib (2010) uses word frequency measured by TF-IDF to enable classification. (2) Features in the experiments are based on estimations provided by human experts. For example, in Ghazizadeh et al. (2008), the fuzzy inference engine uses experts' subjective views on the chain of narrators to determine features. Both of the above-mentioned aspects have significant weaknesses.

By using aspects of words, semantic and real-world similarities are not considered. Existing approaches are not able to classify or extract Hadiths according to their semantics, e.g. by considering the similarity between "quit" and "resign". An example for a Hadith that cannot be classified correctly is given below. The words mentioned in this short Hadith may appear more in the Jihad subject than in the almsgiving subject, while the Hadith itself belongs to almsgiving subject.

"Adi b. Hatim reported that he heard Allah's Messenger (may peace and blessings be upon Him) as saying: He who among you can protect himself against Fire, he should do so, even if it should be with half a date".

While the second aspect compensates for overlooked semantic similarity in the frequency approaches, it is an extremely labor-intensive and time-consuming process. Therefore, we propose a knowledge based classification method to provide a robust cognitive backbone for the data science and linguistic features of the Hadith. The explosion of linked semantic knowledge graphs has drawn the attention of NLP researchers to this topic. A knowledge graph is a labeled and weighted graph that broadens and itemizes the various concepts appear in a set of words. In addition, some hadiths belong to more than one subject, as exemplified in the Hadiths in the books of Sahih al-Bukhari. Thus, the assignment of Hadiths to subjects represents a multi-label classification problem.

Since the reported classification and data mining experiments have been evaluated using different datasets, we re-evaluated them using the same dataset—an Arabic version of Sahih Al-Bukhari consisting of 3150 Hadiths distributed over 14 groups. We only implement those approaches that used aspects of words because of the labor-intensity of other approaches. In the classification section, the approaches presented by Harrag and El-Qawasmeh (2009), and Harrag et al. (2011b) obtained the highest accuracy rate. They employed an artificial neural network to handle the classification problem. Therefore, it is proven that neural networks are the most suitable solution for the Hadith classification problem because our classification problem deals with high dimensional vectors; at the same time, artificial neural networks can work with a large number of variables.

In the data mining section, the approach presented by Harrag et al. (2008) achieved the highest accuracy rate. It utilized cosine similarity technique and built an enriched query. Cosine similarity is suitable for the Hadith mining problem because it is effective when data are sparse. The query expansion boosts the performance of the system because the query will be adapted to the terminology of the Hadith. Lack of designing knowledge based systems is evident in the literature. As a future work, knowledge based systems can be developed using to complement and refine knowledge acquired from all sources. Knowledge graph is a novel way for natural language defining and modeling that can represent discourses. By using knowledge graphs, we can achieve a higher level of semantic understanding. Therefore, we plan to develop a knowledge based Hadith mining and classification to enhance the results with semantic information gathered from a diverse variety of sources.

Acknowledgments The research for this paper was financially supported by the University of Malaya UMRG Grant (RP003C-14HNE).

References

Abuzeina DEM (2011) Utilizing data-driven and knowledge-based techniques to enhance Arabic speech recognition. King Fahd University of Petroleum and Minerals, Saudi Arabia

Aizawa A (2003) An information-theoretic perspective of tf-idf measures. Inf Process Manag 39(1):45–65. doi:10.1016/S0306-4573(02)00021-3

Al Kharashi IA, Al Sughaiyer IA (2002) Rule merging in a rule-based Arabic stemmer. In: Proceedings of the 19th international conference on computational linguistics, vol 1. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 1–7. doi:10.3115/1072228.1072265

Aldhaln K, Zeki A, Zeki A (2010) Datamining and Islamic knowledge extraction: alhadith as a knowledge resources. In: Proceeding 3rd international conference on ICT4M, Jakarta, Indonesia, pp 21–25. Retrieved from http://irep.iium.edu.my/17123/1/WA_17123_AKRAM_Datamining_and_Islamic_KnowledgeExtraction.pdf

Aldhaln K, Zeki A, Zeki A (2012) Knowledge extraction in hadith using data mining technique. Int J Inf Technol Comput Sci 2:13–21. Retrieved from http://www.ijitcs.com/2ndicekmt/Kawther+AAldhaln.php

Aldhaln K, Zeki A, Zeki A, Alreshidi H (2012a) Improving knowledge extraction of Hadith classifier using decision tree algorithm. In: 2012 international conference on information retrieval knowledge management, pp 148–152. doi:10.1109/InfRKM.2012.6205024

Aldhaln K, Zeki A, Zeki A, Alreshidi H (2012b) Novel mechanism to improve Hadith classifier performance. In: 2012 international conference on advanced computer science applications and technologies (ACSAT), pp 512–517. doi:10.1109/ACSAT.2012.93

Al-Kabi MN, Al-Sinjilawi SI (2007) A comparative study of the efficiency of different measures to classify Arabic text. Univ Sharjah J Pure Appl Sci 4(2):13–26

Alkhatib M (2010) Classification of Al-Hadith Al-Shareef using data mining algorithm. In: European, mediterranean and middle eastern conference on information systems, EMCIS2010, Abu Dhabi, UAE, pp 1–23. Retrieved from http://www.iseing.org/emcis/emcis2010/Proceedings/AcceptedRefereedPapers/C20.pdf

Alrazou HM (2004) Computerized frame of the Prophetic tradition. In: 17th national conferences for computer, Medina, Saudi Arabia, pp 596–610

Alrazou HM (2008) Data mining application on the Islamic knowledge resource. Alukah. Retrieved from http://www.alukah.net/culture/0/3123/

Al-tarawneh R, Hamatta HSA, Muiadi H (2014) Novel approach for Arabic spell-checker: based on radix search tree. Int J Comput Appl 95(7):1–5

Althobaiti M, Kruschwitz U, Poesio M (2014) AraNLP: a Java-based Library for the processing of Arabic text. In: Calzolari N (Conference Chair), Choukri K, Declerck T, Loftsson H, Maegaard B, Mariani J, et al (eds) Proceedings of the ninth international conference on language resources and evaluation (LREC'14), European Language Resources Association (ELRA), Reykjavik, Iceland

Attia M, Toral A, Tounsi L, Monachini M, van Genabith J (2010) An automatically built named entity lexicon for Arabic. In: Calzolari N (Conference Chair), Choukri K, Maegaard B, Mariani J, Odijk J, Piperidis S, et al (eds) Proceedings of the seventh international conference on language resources and evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta

Atwell E, Brierley C, Dukes K, Sawalha M, Sharaf A-B (2011) An Artificial intelligence approach to Arabic and Islamic content on the internet. In: Proceedings of NITS 3rd national information technology symposium, The University of Leeds, pp 1–13. doi:10.13140/2.1.2425.9528

Batyrzhan M, Kulzhanova BR, Abzhalov SU, Mukhitdinov RS (2014) Significance of the hadith of the Prophet Muhammad in Kazakh proverbs and sayings. Proced Social Behav Sci 116:4899–4904. doi:10.1016/j.sbspro.2014.01.1046

Bilal K, Mohsin S (2012) Muhadith: a cloud based distributed expert system for classification of ahadith. In: Proceedings of the 2012 10th international conference on frontiers of information technology, IEEE Computer Society, Washington, DC, USA, pp 73–78. doi:10.1109/FIT.2012.22

Cabena P, Hadjinian P, Stadler R, Verhees J, Zanasi A (1998) Discovering data mining: from concept to implementation. Prentice-Hall Inc, Upper Saddle River, NJ, USA

Choi FYY (2000) Advances in domain independent linear text segmentation. In: Proceedings of the 1st North American chapter of the association for computational linguistics conference, Association for Computational Linguistics, Stroudsburg, PA, USA, pp 26–33. Retrieved from http://dl.acm.org/citation.cfm?id=974305.974309

Fattāhizādeh F, Afshāri N (2010) An approach to understanding hadith in Wasāil al-Shia. Hadith Stud 4:67–98

Flachsbart B, Bond WE, St. Clair DC, Holland J (1994) Using the ID3 symbolic classification algorithm to reduce data density. In: Proceedings of the 1994 ACM symposium on applied computing, ACM, New York, NY, USA, pp 292–296. doi:10.1145/326619.326750

Ghazizadeh M, Zahedi MH, Kahani M, Bidgoli BM (2008) Fuzzy expert system in determining hadith validity. In: Sobh T (ed) Advances in computer and information sciences and engineering. Springer, Netherlands, pp 354–359. doi:10.1007/978-1-4020-8741-7_64

Hanum HM, Bakar ZA, Rahman NA, Rosli MM, Musa N (2014) Using topic analysis for querying halal information on Malay documents. Proced Social Behav Sci 121:214–222. doi:10.1016/j.sbspro.2014.01.1122

Harrag F (2014) Text mining approach for knowledge extraction in Sahîh Al-Bukhari. Comput Hum Behav 30:558–566. doi:10.1016/j.chb.2013.06.035

Harrag F, El-Qawasmeh E (2009) Neural network for Arabic text classification. In: Second international conference on the applications of digital information and web technologies, 2009. ICADIWT '09, pp 778–783. doi:10.1109/ICADIWT.2009.5273841

Harrag F, El-Qawasmeh E, Pichappan P (2009) Improving arabic text categorization using decision trees. In: First international conference on networked digital technologies, 2009. NDT '09, pp 110–115. doi:10.1109/NDT.2009.5272214

Harrag F, El-Qawasmeh E, Salman Al-Salman A (2011a) Extracting named entities from prophetic narration texts (hadith). In: Zain J, Wan Mohd W, El-Qawasmeh E (eds) Software engineering and computer systems, vol 180. Springer, Berlin, pp 289–297. doi:10.1007/978-3-642-22191-0_26

Harrag F, El-Qawasmeh E, Salman Al-Salman A (2011b) Stemming as a feature reduction technique for Arabic text categorization. In: 2011 10th international symposium on programming and systems (ISPS), pp. 128–133. doi:10.1109/ISPS.2011.5898874

Harrag F, Hamdi-Cherif A (2007) UML modeling of text mining in Arabic language and application to the prophetic traditions "Hadiths." In: The 1st international sysmposium on computers and Arabic language and exhibition, KACST & SCS, pp 11–20. Retrieved from iscal.org.sa

Harrag F, Hamdi-Cherif A, El-Qawasmeh E (2008) Vector space model for Arabic information retrieval—application to "Hadith" indexing. In: First international conference on the applications of digital information and web technologies, 2008. ICADIWT 2008, pp 107–112. doi:10.1109/ICADIWT.2008.4664328

Harrag F, Hamdi-Cherif A, Salman Al-Salman A, El-Qawasmeh E (2009) Experiments in improvement of Arabic information retrieval. In: 3rd IEEE international conference on Arabic language processing, Rabat, Morocco, pp 71–81

Hsu C-W, Lin C-J (2002) A comparison of methods for multiclass support vector machines. IEEE Trans Neural Netw 13(2):415–425. doi:10.1109/72.991427

Hu K, Lu Y, Zhou L, Shi C (1999) Integrating classification and association rule mining: a concept lattice framework. In: Zhong N, Skowron A, Ohsuga S (eds) New directions in rough sets, data mining, and granular-soft computing, vol 1711. Springer, Berlin, pp 443–447. doi:10.1007/978-3-540-48061-7_53

Jbara K (2010) Knowledge discovery in Al-Hadith Using text classification algorithm. J Am Sci 6(11):485–494

Kabi MNA, Kanaan G, Al-Shalabi R, Al-Sinjilawi SI, Al-Mustafa RS (2005) Al-Hadith text classifier. J Appl Sci 5(3):584–587. doi:10.3923/jas.2005.584.587

Kamsin A, Gani A, Suliaman I, Jaafar S, Mahmud R, Sabri AQM, et al (2014) Developing the novel Quran and Hadith authentication system. In: 2nd international conference on islamic applications in computer science and technology, Amman, Jordan, pp 1–8. Retrieved from http://umexpert.um.edu.my/file/publication/00006192_111415.pdf

Karim NSA, Hazmi NR (2005) Assessing islamic information quality on the internet: a case of information about hadith. Malays J Libr Inf Sci 10(2):51–66

Lacroix S (2008) Al-Albani's revolutionary approach to hadith. ISIM Rev 21(1):6–7

Maazouzi F, Bahi H (2012) Using multi decision tree technique to improving decision tree classifier. Int J Bus Intell Data Min 7(4):274–287. doi:10.1504/IJBIDM.2012.051712

Melchert C (2002) Early renunciants as Hadīth transmitters. Muslim World 92(3–4):407–418. doi:10.1111/j.1478-1913.2002.tb03750.x

Najeeb MM (2014) Towards innovative system for Hadith Isnad processing. Int J Comput Trends Technol 18(6):257–259. doi:10.14445/22312803/IJCTT-V18P154

Papadimitriou CH, Tamaki H, Raghavan P, Vempala S (1998) Latent semantic indexing: a probabilistic analysis. In: Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on principles of database systems, ACM, New York, NY, USA, pp. 159–168. doi:10.1145/275487.275505

Ragas H, Koster CHA (1998) Four text classification algorithms compared on a Dutch corpus. In: Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval, ACM, New York, NY, USA, pp 369–370. doi:10.1145/290941.291059

Rennie JD, Shih L, Teevan J, Karger D (2003) Tackling the poor assumptions of naive bayes text classifiers. In: Proceedings of the twentieth international conference on machine learning, ICML-2003, Washington DC, pp 616–623

Roche E, Shabes Y (eds) (1997) Finite-state language processing. MIT Press, Cambridge, MA, USA

Salton G, Wong A, Yang CS (1975) A vector space model for automatic indexing. Commun ACM 18(11):613–620. doi:10.1145/361219.361220

Sebastiani F (2002) Machine learning in automated text categorization. ACM Comput Surv 34(1):1–47. doi:10.1145/505282.505283

Shatnawi MQ, Abuein QQ, Darwish O (2012) Verification hadith correctness in islamic web pages using information retrieval techniques. Int J Comput Appl. doi:10.5120/6327-8680

Tata S, Patel JM (2007) Estimating the selectivity of tf-idf based cosine similarity predicates. SIGMOD Rec 36(4):75–80. doi:10.1145/1361348.1361351

The Encyclopedia of the Nine Books for the Honorable Prophetic Traditions! (1997). Sakhr Company. Retrieved from http://www.harf.com

Yang Y, Liu X (1999) A re-examination of text categorization methods. In: Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval, ACM, New York, NY, USA, pp 42–49. doi:10.1145/312624.312647

Yusoff Y, Ismail R, Hassan Z (2010) Adopting hadith verification techniques in to digital evidence authentication. J Comput Sci 6(6):613–618

Zhang M-L, Zhou Z-H (2005) A k-nearest neighbour based algorithm for multi-label classification. In: 2005 IEEE international conference on granular computing, vol. 2, pp 718–721. doi:10.1109/GRC.2005.1547385