Predicting Semantic Categories in Text Based on Knowledge Graph Combined with Machine Learning Techniques

Mohamed Atef Mosa

Published online: 19 Aug 2021.

Submit your article to this journal ☑

Article views: 900

View related articles ☑

View Crossmark data ☑

Citing articles: 5 View citing articles ☑

Taylor & Francis
Taylor & Francis Group

Check for updates

Predicting Semantic Categories in Text Based on Knowledge Graph Combined with Machine Learning Techniques

Mohamed Atef Mosa (iD)

Information Technology Sector, Institute of Public Administration, Riyadh, KSA

ABSTRACT

The Quran and the Sunnah are the two principal elements of the Islamic religion, and the hadith is an interpreter of the Quran. Hadith is everything that the Messenger Muhammad said, whether it was a word, an action, or a good adjective of the Prophet. Given the status of the hadith of Muslims everywhere in the world, digging into it is the m ain perspective to evoke the guiding principles and institutions that Muslims must follow. The mining of the hadith has received much attention in recent times, but so far, the work has not been fully implemented. This study focuses on predicting the semantic categories of the unclassified hadith text based on its text. The model can distinguish between several categories to predict the optimal one such as ablution, fasting, Hajj, and Zakat. To achieve this goal, a Knowledge-Graphic (KG) prediction model was developed to improve machine learning classifiers from the standpoint of two unique traits. 1) Define pivotal terms that have high values. II) Taking into account all the paths of those pivotal terms through their convergence with the categories in the Knowledge-Graphic of all the paths that link them. We rely on six books with more than 30,000 hadith and 120 classifications. Empirically, we found optimistic results in combining the KG model and machine learning classifiers.

Introduction

The two prime sources of Islamic guidance stem from the Quran and hadith. The Noble Quran is the essential source in Islam that includes the words of God as revealed to the Prophet Muhammad. Also, the hadith are the complementary source that unambiguously interprets some verses of the Holy Quran. For Muslims to live their days as God commanded them, they must abide by the teachings of Islam, and they must follow the established rules. Digital transformation has allowed the Quran and Sunnah to be accessed digitally and disseminated to everyone at all times to enable Muslims to research precisely the topic they need (Rostam and Malim 2019). The Hadith classification has

CONTACT Mohamed Atef Mosa ✉ mosamo@ipa.edu.sa mohammedatefmosa@gmail.com; Information Technology Sector, Institute of Public Administration, Riyadh, KSA

become one of the most important topics to be addressed with care. That may be useful for the Mufti "someone who pries through the Quran and Sunnah to extrapolates instructions and rules to give to the public."

Text classification refers to the method for describing untagged text documents in at least one pre-defined category based on their content (Adeleke et al. 2017; Bakar et al. 2018). A few text classifications cases have been developed, for example, computation of K-Neighbor Nearest (KNN), Naive Bayes (NB), Supporting Vector Machines (SVM), and Neural Organizations (NNW). In addition, some hybrid or group research methods have been developed to improve basic classifiers such as Adaboost and Bagging (Elghazel et al. 2016; Hassanat et al. 2014). Text categorization has been broadly applied in text-based datasets, such as essays (Dong and Zhang 2016). Where 423 articles were examined and peer-reviewed, 12,881 references, 101019 sentences published in the library, and information science literature in Turkey meticulously. Citations are divided into four main categories. The tagging process was undertaken with the agreement of commenters and citation categories defined for quote sentences. The auto-quote sentence ranking has a 90% success rate. Additionally, Facebook and Twitter status updates are the hot applications of classification and other domains (Mosa, M. A. et al. 2017a, Mosa, M. A. et al. 2017b, Mosa, M. A. 2017c, Mosa, M. A. 2019a, Mosa, M. A. et al. 2019b, Mosa, M. A. 2020a, , Mosa, M. A. 2020b). Where thousands of comments on social media are summarized using Ant Colony Optimization, Gravitational Search Algorithm, and Particle Swarm Algorithm as single/multi-objective tasks. The task of categorizing hadith is one of the most noteworthy applications that enhanced the accessibility of hadith based on topics. Nevertheless, for the field of Islamic purse, most studies seem to focus exceedingly on ontology. Most researchers developed Quran ontology for classification depending on themes to define a searching method using semantic relations (Suryana, Utomo, and Mohd Sanusi 2018). Meanwhile, other studies focused on frequent pattern methods to get forth association rules (Zainol et al. 2016).

In this paper, our goal is to predict the semantic category of unclassified hadith based on a cognitive graph in combination with machine learning classifiers. In the Sunnah of the Prophet, the same hadith may be classified under several categories. In this case, the model predicts the hadith's category to more than one based on the highest scores. On the other hand, the classification of the prediction task can be done by a specialist of Hadith scholars where the categories are required to evaluate the quality of the classification. The human evaluation may sometimes lead to evaluation mistakes due to the multiplicity of classification for a single hadith, and the hadith text's lack of class-specific semantics terms. Alternatively, we can adopt an evaluation model that depends entirely on the machine as shown in the

experimental analysis section. The prediction model will be based on a training set and then used to predict items for a categorized test set. In the testing phase, ratings are removed for the testing process to be accurate.

The remainder of the paper is structured as follows: Section 2 presents related work. In Section 3, the proposed approach to predicting semantic categories is presented. Section 4 exhibits experimental analysis and results. Section 5 provides conclusions.

Related Work

In (Kamsin et al. 2014), the authors affirmed the importance of having an automated authentication system for the Quran and the hadith of the Prophet for confrontation rigging copies of the Quran and hadith in the virtual domain. In (AL-Kabi, Mohammed N., et al. 2015), the authors expound that the hadiths classification task can be dealt with simply. They classified the Arabic hadith texts into eight categories of Sahih al-Bukhari by counting the repetition of the term. After driving out stop-words and Isnad, each word was transformed into its stem using a stemmer system. In Sahih al-Bukhari, the same hadith may be appended to more than one category. In this case, the system offered two subjects of the highest ranks by performing term weighting based on the NLP methods such as TF-IDF and trained by 120 Hadiths. The system relied on a fresh eighty Hadiths for testing. The system has eventually brought about an accuracy of 83.2%.

Additionally, (Ghazizadeh et al. 2008) developed a system based on the fuzzy expert by collecting rules and skillful opinions. To conclude, two heuristics engines are designed in the consecutive form. The first engine brings forth the rank of the narrator and passes it to the second one. The second stage output is the speech validation rate. The system had been tested by labeling the unknown type of hadith by weak, good, reliable, and correct degrees to their reliability rating. The system nailed down up to 94% accuracy through a combination of triangular fuzzier, single fuzzier, multiplication fuzzier, and average clearance of de-fuzzier.

On the other hand, several methods have been developed based on machine learning techniques. Artificial Neural Networks (ANN) is one of the most popular classification methods. In (Harrag, El-Qawasmeh, and Pichappan 2009), the authors used the ANN method for classifying hadiths. In a classification approach, the pre-processing phase is performed initially. There are more than 700 unique vocabularies in the dataset, each feature indicated a single term. Finally, the sparse and high-dimensional vector converts into a vector 200 dimensions of the document term weights. Some systems are developed based on a decision tree classifier. In (Maazouzi and Bahi 2012), leaves refer to a class, and labels denote the coincidence of attributes that refer to specific labels. Additionally, for traversing across the tree, a top-down

algorithm is developed to predict classes. In (Harrag and El-Qawasmah 2009), authors developed a classification model by experiment with more than 400 hadith distributed among 14 groups of Encyclopedia of the Prophet via ID3 classifier (Flachsbart et al. 1994). The pre-processing stage consists of transforming the document into soft text, removing stop-words, and stemming. After pre-processing, a vector was created consisting of all the terms in the text of the hadith. Then, the dimension of the vector was reduced to 1938 based on some specific criteria. Next, a weight for each dimension was calculated using the term frequency (TF). The evaluation in the test phase resulted in a recall of 38%, an accuracy of 47%, and a score of 40%. There were many errors in the classification due to the nature and characteristics of the hadith documents. Several classes have been labeled such as Hajj, Prayer, and Zakat. Three classifiers have been employed to fulfill the prediction task based on Naïve Bayes (NB), Support Vector Machine (SVM), and K-Nearest Neighbor (KNN). Besides, the SVM brought about better accuracy with an improvement of 10% −20%, compared to the other methods which were able to show little improvement in terms of accuracy (Rostam and Malim 2019). Finally, several traditional supervised learning algorithms with alternative combinations to enhance classification accuracy was proposed (Abdelaal, Elemary, and Youness 2019). Three classifiers, decision tree (DT), NB, and random forest (RF) were primarily judged, and the results were promising. Our proposed algorithm is developed in combination with various machine learning techniques to improve the accuracy of the overall system.

Proposed Model

Classification of the hadith is the process of predicting the category of untitled hadith according to a predefined set of titles. The model is developed based on analyzing training data whose categories are known in advance. Taking into consideration that it is permissible for scholars to classify the hadith itself into more than one category because its material belongs to them all. The categories in which the hadith are classified are many, including monotheism, fasting, prayer, zakat, Hajj, and the like. In this section, the specifics of the hadith classification task are decomposed and explained in detail. The overall process of the approach unfolds beyond its parts in the Figure 1. First, collect the texts of the hadith from the authentic Sunnah sources. The Six Books of Bukhari, Muslim, etc., are the books on which the model relied on the training and testing phases. Later, some natural language processing techniques are applied to prepare and clean the text. In addition, the method of stemming is employed to unify family words that appear in different shapes into a single shape. TF-IDF and Mutual Information (MI) techniques are developed for pivot selecting terms that have semantic meaning and are frequently repeated. Then, a knowledge graph is constructed based on these notable words
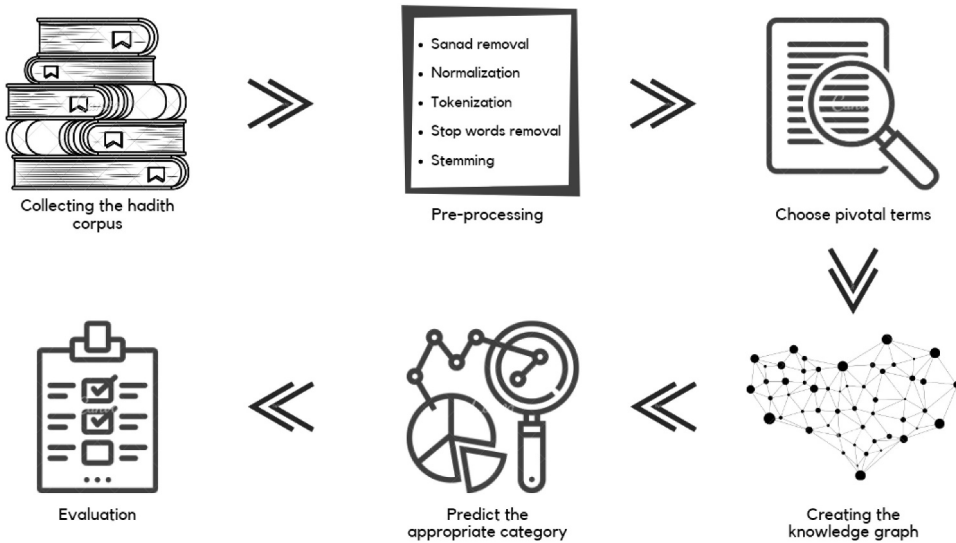
Figure 1. Hadith categorization model.

appearing concurrently with the reference categories together in the same text. The main aim behind the knowledge graph is that the link between two nodes holds the value of how closely the two words are related to each other, whether the term is a category or a word belonging to the text. The proposed knowledge graph model is constructed to precede machine learning classifiers for the eventual prediction phase.

Collecting the Hadith Corpus

The hadith is compiled from the most popular six books of Sunnah containing Sahih Al-Bukhari and Sahih Muslim. In this study, the number of categories/ topics used in the training process is more than 120 categories/topics. Each Hadith belongs to mostly a specific topic or sometimes more than one topic. The dataset is split into two main sections. The first one is assigned for training, and the second one is for testing sets using the cross-validation method. All corpus is divided randomly into several n-blocks. Each is suspended once, and the classifier is trained on the remaining (n-1) fragments. The cross-validation method is superior for appraising the accomplishment of the model.

Pre-processing

The hadith of the Prophet contains two main parts: the text (which is the content of the hadith) and the chain of a person's name (series of narrators of hadith from the Companions and Taabi'een) (Saloot et al. 2016). In this study,

we do not need hadith Sanad, it is an abandoned part since it does not contain significant semantics of the text. The Narrators series is more concerned with classifying hadith in terms of the degree of authenticity of hadith. The most crucial step in the classifying task is the NLP phase. The topic of separating the hadith text from its text has been one of the topics that researchers have been interested in recently. This point is considered one of the essential points on which the accuracy of classification depends. Although there is little research on this point, the authors (Maraoui, Haddar, and Romary 2018) have adopted a method to separate the hadith from its text. The study relied on the book Sahih Al-Bukhari in the mechanism of its work. The results obtained were encouraging despite some drawbacks related to exceptional cases due to the structure of the hadith. In this thesis, the transcripts were separated from the islands in the data set, so this step was overlooked.

Second, the terminology normalization process aims to unify the shapes of some letters that appear in several different forms in the same term into one form. Third, Tokenization that aims to divide the text into separated terms by spaces like white space, semicolons, and commas. Fourth: Stop word removal that aims to remove unimportant words, such as those that appear in sentences frequently and have no semantic meaning or effect on the content, like "in في" "onيلع" "toيلإ" "..., etc. (El-Khair 2006). Finally, the stemming or lemmatization task, where the stemming technique aims to remove all suffixes such as prefixes, suffixes, and suffixes from the word to diminish the emergence of different forms of the term that reflect the equivalent sense. And lemmatization means root extraction task for a word's family, where all of these words share the same abstract meaning. Generally, stemming or lemmatization improves prediction and classification accuracy (Hajjar et al. 2010).

Selecting Pivot Terms

The main objective behind this section is to determine the most frequently pivot terms in the hadith text. Considering all of them for relevant category prediction may be almost confuse the model. Considering and relying on all terms that contain accidental and repetitive terms alike in the classification process may be imprecise, as words that do not arise in the text cannot be relied on permanently. Therefore, proposing a method for choosing a major set of terms from a corpus based on their high informational content is vital and effective, so that these terms are the keywords for every hadith. While there are several ways to define the pivotal terms, we consider two methods – by term frequency-inverse document frequency (TF-IDF) and mutual information (MI) based on the entropy.

For choosing pivotal terms using TF-IDF as shown in equation NO. 1, where is composed of two parts: $tf(w_i)$ is the number of times a term w_i appears in the list of Hadith. And $idf(w_i)$ is the logarithm of the total number of the Hadiths divided by the number of Hadiths that the term w_i appears in.

$$tf - idf(W_i) = \sum_{w_i C_i} tf(W_i).logidf(W_i) \tag{1}$$

In terms of entropy-based pivot term election, we eliminated terms with great entropy (which tend not to be specific, they are general and non-informative terms) and identify terms with low entropy (which tend to be more specific). The entropy of term t is measured as shown in equation NO. 2.

$$Entropy(t) = -\sum_{h_i H} p(h_i|t) * lo(p(h_i|t)) \tag{2}$$

Where H is the list of all Hadith that co-occur with term i. $p(h_i|t)$ is the probability of appearing a term t in Hadith h_i. By selecting as pivot terms those terms are low in entropy. Generally, the goal is to find good predictors of categories.

Creating the Knowledge Graph

In this section, we suggest a graph-based knowledge prediction approach. The basic idea is to define the implicit relations among the categories and terms used in Hadith to build a knowledge diagram. As a consequence, it can then be used to link terms in unlabeled Hadith texts to the appropriate categories. The basic assumption is that the terms and categories used together are related to each other and thus close in meaning. First, we build a knowledge graph and suggest a path assembly technique to record the convergence between terms and categories in the proposed graph. As illustrated in Figure 2, we build a graph G = (N, E) with nodes N = {n1, n2 ... n$_m$} in which n$_i$ is either a term or a category and edges E = {e1, e2 ... e$_r$} in which e$_j$ is the weighted edge between two nodes. To avoid confusion and keep the graph less contaminated, an edge of two nodes is simply created when the number of co-occurrences is not less than a predefined threshold. The common occurrence between terms and each other is measured by looking at all the hadiths in the training corpus and counting the number of times the two terms (either terms or categories) appear together in the same hadith text. This way, we may filter out any edges that occurred by chance. But what is the proper weighting function for edges between terms (nodes)? This weighting function can be used to determine the relative convergence of terms and categories together that are directly related. For instance, if we consider the case where the terms A and B appear 100 times together, and both of them appear in the corpus exactly 60
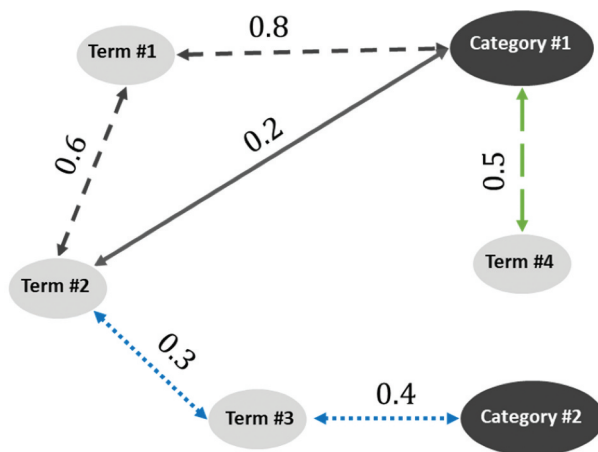
Figure 2. Hadith knowledge graph.

times. In this case, the terms A and B always occur together and never separate. Now suppose the term A appears 90 times, but only 20 times with B. In this case, the convergence between A and B is less than it was in the first case. Therefore, we normalize the value of repetition by counting the number of times a term appears in the whole corpus which is equal to the number of external links for that node. This normalized weight score, $NT(n, n + 1)$ is the normalized weight of the edge between node n and node $n + 1$:

$$NT(n, n + 1) = \frac{NT(n, n + 1)}{\sum_{p \in outlink(n)} NT(n, p)} \qquad (3)$$

Where $NT(n, \ n + 1)$ is the co-occurrence count of the two terms n and $n + 1$ (terms of Hadith or categories). $NT(n, p)$ is the sum of the numbers of occurrence of the term n with words other than the term n +1. Through this normalization, we consider how much one node devotes to the relationship with another node. Hence, the edge to a more general term will have a smaller weight compared to the feature of a more specific term.

The Weighting of Terms

According to the previous normalization, we consider how much the two nodes are bounded. The question now is how to measure the convergence of nodes that are more than one jump away (i.e., that has more than one node to reach the category)? What would happen if the relevant category was found in two or three jumps? And to what level? When calculating the extent of convergence between the term and category, all paths plotted in the graph must be taken into consideration. At the same time, the scores for categories that are further away from a particular term should be punished? To formalize

the problem, we assume that category C is reachable from term $\mathbf{t_i}$ in radius j as $\mathbf{t_i}\, j\, \mathbf{C}$. For a separate path of a term to a category, by considering all the weight of edges in the path, all values are multiplied, where the edge weights themselves are come down by a decay factor α. The main aim behind the decay factor is to punish the terms farther from the category so that we still consider them to be candidate categories but of less importance than those directly related in the graph. The score of the path, then, is:

$$\mathbf{score\_path}\left(\mathbf{t_{i\rightarrow}^{j}},\mathbf{C}\right) = \prod_{\mathbf{n}=0}^{\mathbf{m}-1} \mathbf{NT}(\mathbf{n},\mathbf{n}+1)*\alpha^{*} \qquad (4)$$

Where the normalized edge weights are conducted between terms, the further the term is from the category node, the smaller the score it will get. To find the total score of category C from the term $\mathbf{t_i}$, we measure the aggregation scores for all paths in between. Therefore, if a category can be obtained through several paths, it shows higher relevance to the term compared to one that can only be obtained by one path. Hence, the total of these path points is:

$$\mathbf{term\_score}(\mathbf{t_i},\mathbf{C}) = \sum_{\mathbf{m}=1}^{\mathbf{M}} \mathbf{score\_path}\left(\mathbf{t_{i\rightarrow}^{m}}\mathbf{C}\right) \qquad (5)$$

All paths from a term to a category are considered. For instance, in Figure 2. category $Category\#1$ is reachable from four terms $term_1, term_2, term_3, term_4$. Where each term is connected to the category through a path of a different color and shape. On the other hand, $term_1$, and $term_4$ reach to the category in one jump, $term_2$ through two and one jumps (from two different paths), and finally, t_3 through three jumps. In this way, we link terms to categories considering all potential paths.

Eventually, the relationship among terms of Hadith and categories is conducted based on the PageRank algorithm. The hypothesis behind that is to highlight the strong links among terms and their categories. For instance, when there is a term T_1 votes (i.e., belongs) to just one category C_1, and term T_2 votes to categories C_1, C_2, and C_3. The linkage between T_1 and C_1 is stronger than the linkage between T_2 and C_1. The PageRank algorithm is reformed to fulfill this issue. To calculate the term-rank weights, a random walk is performed among terms and categories. The more the term t_i appears with several categories C, if it belongs to more than one category, the less the linkage between the term t_i and the categories C associated with. Therefore, at accounting score between term t_i and category C, it will be divided by the number of occurrence term t_i with other categories associated with it. That

means associating a term with only one category will make the linkage between them equal one. The equation for calculating the term rank is shown below.

$$term\_score(t_i, C_i) = \alpha * \frac{term\_score(t_i, C_i)}{\sum_{C_j L_{C_j}} term\_score(t_i, C_j)} + (1 - \alpha) \tag{6}$$

Where $term\_score(t_i, C_i)$ is the weight of linkage between term t_i and category C_i, L_{C_j} is the list of weights of out-links from term t_i to other categories C_j. The initial value of all terms is 1, random walk iterations are going on until stabilization of results and no changing them. The damping factor is set at 0.85 according to (El-Fishawy, N. et al. 2014). Eventually, we will obtain the actual weights for just the pivot terms, as exhibited in Table 1: The value varies from one category to another.

Knowledge-graph Based Machine Learning Classifiers

Knowledge-graph is an innovative way instead of the feature selection stage for solving the high dimensional data problem by removing redundant and unwanted data to improve the classification task (Bahassine et al. 2020). Knowledge-graph is employed to weight and select only the pivot terms and scored as mentioned in Table 1. The data in Table 1 is as input for several machine learning classifiers such as NB, DT, and RF as illustrated in (Abdelaal, Elemary, and Youness 2019).

Evaluation Analysis

In this section, an experimental study of the Hadith category recommendation is got on with. We introduce the dataset and metrics used and present several alternative modifications to predict the Hadith category, and introduce the results of a comparative study.

Table 1. The weighted pivot terms in Hadith corpus.

Hadith	Term #1	Term #2	...	Term #m	Category
H #1	0.23	0.53	...	0.43	**C #1**
H #2	...	0.53	...	0.43	**C #1**
H #3	0.23	...	...	0.43	**C #2**
	...	...	...	...	**...**
H #n	0.33	...	...	0.35	**C #f**

Dataset

For the experimental evaluation, we compiled Hadith text from six books. Sahih Al-Bukhari, Sahih Muslim, Sunan Ibn Majah, Sunan Al-Nasa'I, Al-Jami 'by Al-Tirmidhi, and Sunan Abi Dawood. The corpus contains 39,038 annotated Hadith that comprises more than 10 million tokens and about 130 categories/topics. Each Hadith mostly belongs to one topic or sometimes more than one topic. The dataset is actually sectioned into two main parts: the first one is represented in 80% as a training set, the rest is for the testing phase. The cross-validation method is employed by dividing the corpus into five separated sections randomly. All categories are represented in all sections in proportion to their percentage in the training data. One section is suspended for each run, and the classifier is trained on the remaining sections.

Experimental Results

In general, a framework is adopted for the evaluation in terms of dividing the data into a training group for training the model, and a separate test group for the evaluation. The model is used to predict categories of test Hadiths by removing all categories. Expected C categories are referred to as $\mathbf{C_{predicted}}$. Where $\mathbf{C_{predicted}}$ means true-positives (TP) + false-positive (FP). The actual categories applied to the hadiths are $\mathbf{C_{real}}$. Where $\mathbf{C_{real}}$ means true-positives (TP) + false-negative (FN). For different C-values, we can evaluate the quality of category prediction using precision:

$$\mathbf{precession} = \frac{\left|\mathbf{C_{real}} \cap \mathbf{C_{predicted}}\right|}{\left|\mathbf{C_{predicted}}\right|} \tag{7}$$

Where the prediction of actual categories used results in the only accuracy of 1 while predicting that none of the correct categories used results in an accuracy of 0. Besides, we evaluate the quality of the category's prediction using recall:

$$\mathbf{recall} = \frac{\left|\mathbf{C_{real}} \cap \mathbf{C_{predicted}}\right|}{\left|\mathbf{C_{real}}\right|} \tag{8}$$

Where identifying all of the correct category's results in a recall of 1, finally, we also consider the F-measure to make a trade-off between precision and recall:

$$\mathbf{F-measure} = \frac{2*\mathbf{precession} * \mathbf{recall}}{\mathbf{precession} + \mathbf{recall}} \tag{9}$$

We are now judging the achievement of the proposed graph-based approach for categorization. In particular, three categories of experiments are conducted and listed: 1) estimate the parameters used in the graph-based approach and fitting selection methods; (2) Determine the optimal number of categories for the forecasting task. And ultimately, (3) Compare the proposed model with the predefined systems in this section.

Parameter Estimation

We estimate two parameters used by the graph-based approach: 1) the number of gaps considered from term to category. 2) the value of decay factor β that is responsible for punishing terms further from their categories. We also compare two methods for choosing the pivot terms MI and TF-IDF.

In this experiment, we estimate the maximum number of gaps that can be taken from a pivot term to define candidate categories for classification. For instance, since there is a singular gap, this intends that the term is a direct neighbor of the category and rises beside it in the same text. As a count of gaps is greater than one, that means categories do not come up directly with the terms, but they come up with the neighbors of those terms. Hence, in this experiment, we tried with a different number of gaps after setting $\alpha = 0.7$. The result is exhibited in Figure 3. We observed a significant improvement in classification quality when the number of gaps increased to two and three, indicating that these proximal gaps are a good candidate (even if they are not directly adjacent to the categories). We also note that the recall and F-measure increased at gaps 2 and 3, but became decreasing as the number of gaps increased to 4. Therefore, the optimum number of gaps chosen is two, taking into account the model's runtime.
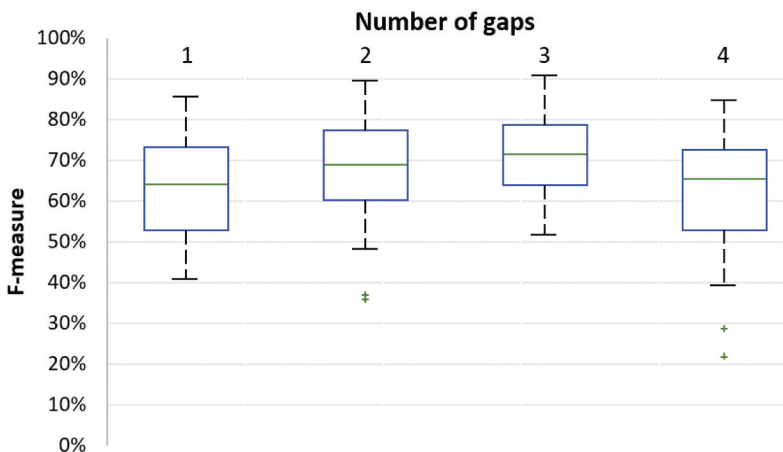

Figure 3. Relevant categories in the knowledge graph based on number of gaps.

Regarding decay factor (α), to determine its near-optimal value and to punish the terms farther from the category, the value fluctuated from 0 to 1, and the performance of the model is recorded. During the experiment, the number of gaps was two. Figure 4 shows the result of all thresholds of the decay factor (α) to calculate the term score as shown in Equation (6). And to balance between the effectiveness of the model and the efficiency, the decay factor (α) is determined to be 0.5.

As illustrated in Section 3.3, a critical problem with the graph-based forecasting approach is determining the proportion of optimal pivot terms that will participate in graph construction. We now assess the performance of our approach using two methods TF-IDF and MI. We got the unique terms of the corpus that have achieved the highest scores based on TF-IDF and MI and took different proportions of them to select the appropriate ratio. Accordingly, we have adopted the highest 3%, 5%, 10%, 15%, and 20% of terms to determine the proportion of optimal pivot terms. We conduct experiments to see which percentage will be most efficient later during the Knowledge Graph building process. The results are exhibited in Figure 5. Interestingly, we picked up that there is a slight variation between the performances of the two approaches. Notwithstanding, the TF-IDF approach is picked for all the subsequent investigations. Ultimately, the percentage of unique terms that participate in building the knowledge graph is 10%.

Analysis Number of Categories

When training the model to predict categories, we mentioned that some Hadiths may belong to more than one category. Therefore, we want to know how the model performs when predicting one class, two, and three classes. We can notice that the model works well, and the performance is high when
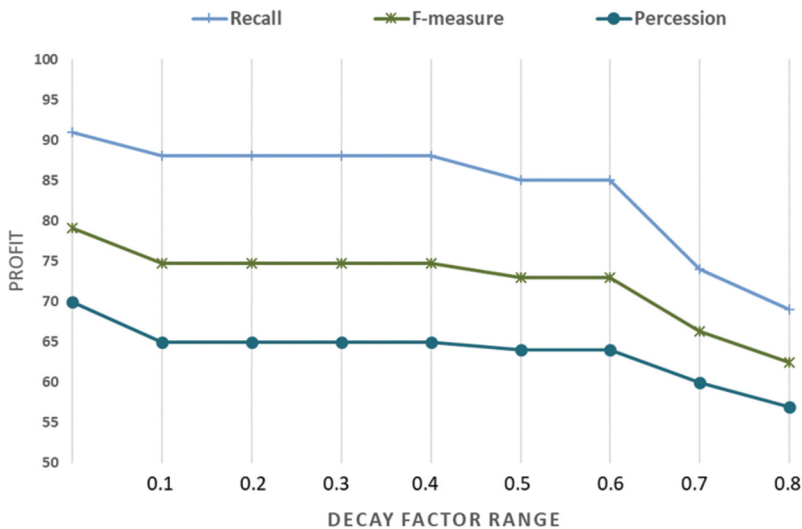

Figure 4. Relevant categories in the knowledge graph based on decay factor (*a*)change.
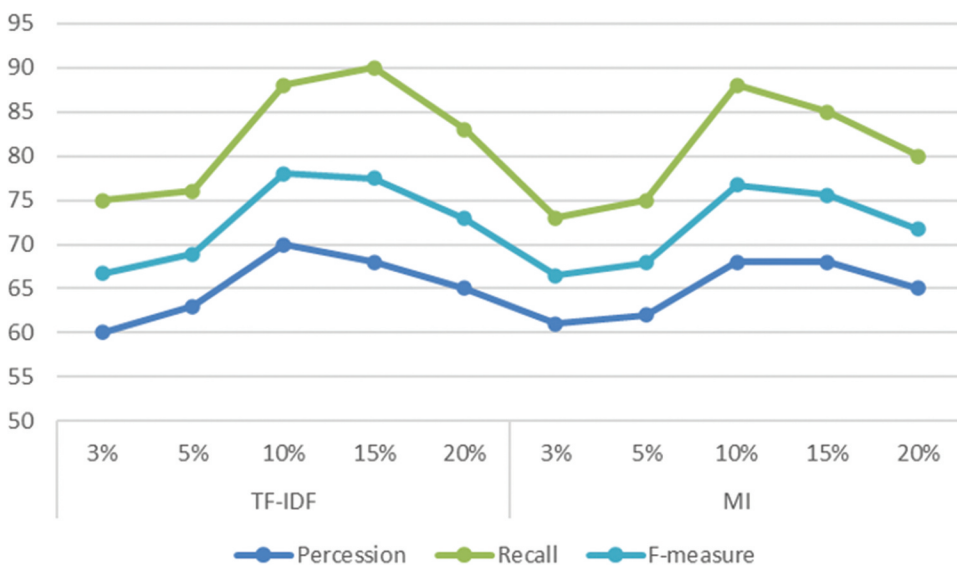
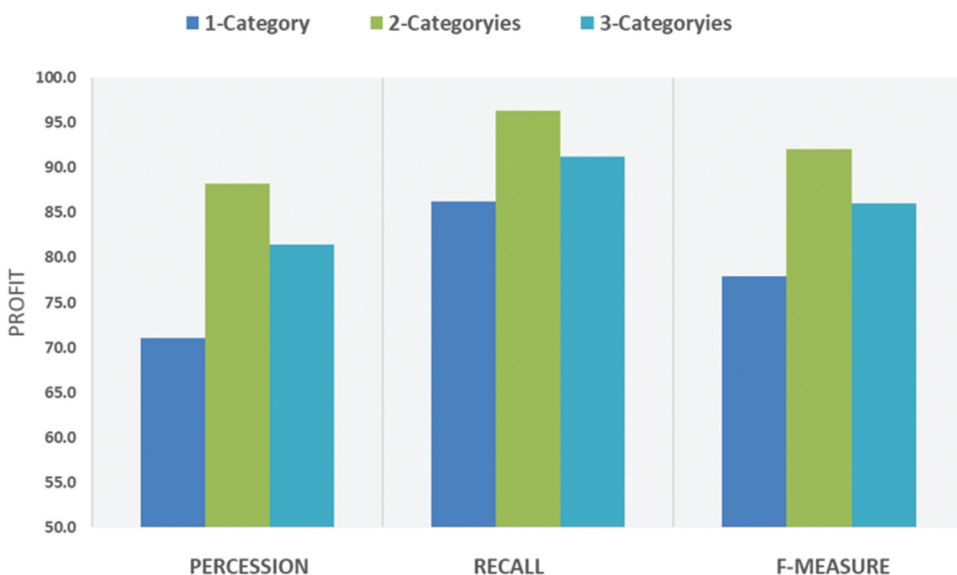**Figure 5.** TF-IDF and MI based pivot terms selection.

Figure 6. Performance against number of categories predicted.

predicting only two categories as shown in Figure 6. If the Hadith is classified into several different categories, then we consider three categories as the maximum for comparison. In the evaluation stage, all hadiths are ranked according to their greatest score. If a hadith is classified only into one or two categories, the outcome is considered based on the highest likelihood. Form

performance has been found to work most out when Hadith is categorized into only two categories ordered. This is because most hadiths classified in more than one category are classified into two categories only in most cases.

Alternative Methods

DT, RF, and NB (Abdelaal, Elemary, and Youness 2019) were the methods that were combined with the proposed model to see their effect on these methods in their conventional form. Initially, the parameters of the algorithms mentioned used as brought up in the research (Abdelaal, Elemary, and Youness 2019). All algorithms predict two categories ranked based on the probability for all Hadiths in the test data. If the Hadith is categorized into only one class, then the first category is considered the outcome of the model. In (Abdelaal, Elemary, and Youness 2019), the Chi-Square and Information Gain models are employed as feature selection for classifiers.

Moreover, we compared our knowledge-graph model with the association rule algorithm. In the Market basket template, we have a large option of items and baskets containing a subset of them (Agrawal, Imieliński, and Swami 1993). We are interested in characterizing considerably purchased substances jointly in a basket. The model generates the association rule of the form $\{t_1, t_2, \ldots, t_n\} \rightarrow \{C\}$, that meaning these substances are gathered with each other in the same basket. Therefore, there is a good chance to find out the appropriate category C with them. In particular, the collective correlation mining approach can be used to define interesting relationships between substances and categories based on the likelihood of substances occurring with related categories. Adapting the market basket model to the Hadith category prediction problem, baskets are a set of Hadith, and items are the terms and categories that appear in a Hadith. The goal is to find out the most proper terms accompanying the categories. In this model, we are concerned with the term pairs that frequently appear together and are highly support values. Confidence is another metric that implies the probability of finding C knowing that it has occurred. High confidence and support mean beneficial association rules. Here we define support (t) as the number of Hadiths in which t has appeared and confidence $\{t\} \rightarrow \{C\}$ as the probability of using categories C when t is observed in a Hadith as a set of terms:

$$confidence(t \rightarrow C) = \frac{support(t, C)}{support(t)} \qquad (10)$$

Which is the count of terms and categories C appear jointly partitioned by the plenty of times that the terms t arose in the training dataset? In this way, association rules are employed to capture attractive term-category relationships. Practically, the most attractive rules have a length of less than three for

short text dataset. Therefore, all possible association rules extracted first from the training set, keep only those of length three or less. To predict the categories for a new Hadith, the rules with the same input of terms and more confidence and support are used.

To evaluate the performance of the ML classifiers in combination with and without the proposed knowledge graph KG model, we need to know how skillful the KG expert is to maximize the accuracy of the rating by identifying the number of properly hadiths categorized and incorrectly categorized hadiths. We used a confusion matrix (CM) to find out the number of hadiths belonging to TP, TN, FP, or FN according to Equations 7, 8, and 9. Tables 3 and 4 show the accuracy of the ML classifiers in the case of combination with/ without the KG model.

In Table 2 We evaluated the performance of the classifiers without the pre-processing stage and the KG model, and it is clear from Table 2 that the results fluctuated between 87.8% to 90.3% for different classifiers with superiority of RF. Table 3 shows a significant improvement when we applied pre-treatment to the data to raise the results from 87.8% to 90.3% and from 91.4% to 93.5%. Finally, by observing Table 4, we can notice the accuracy the classifiers achieved when combined with the proposed KG model. We note that there is a marked improvement in the accuracy of the results and that the cognitive graph was able to significantly improve the results of the classifiers to reach

Table 2. Comparison of categorization accuracy without pre-processing and knowledge-graph.

	Precision	Recall	F-measure
RF	87.20%	96.10%	91.43%
DT	84.10%	92.20%	87.96%
NB	81.50%	94.20%	87.39%
AR	80.90%	95.90%	87.76%

Table 3. Classification accuracy with pre-processing phase.

	Precision	Recall	F-measure
PP-RF	89.50%	97.90%	93.51%
PP-DT	86.50%	96.20%	91.09%
PP-NB	86.00%	95.30%	90.41%
PP-AR	84.50%	96.90%	90.28%

Table 4. Classification accuracy with pre-processing and knowledge-graph phases.

	Precision	Recall	F-measure
PP-KG-RF	93.20%	99.10%	96.06%
PP-KG-DT	90.20%	97.90%	93.89%
PP-KG-NB	89.30%	97.00%	92.99%
PP-AR	84.50%	96.90%	90.28%

96.1%. We obtained a remarkable superiority in the accuracy of the proposed model in precision, recall, and f-measure alike. Knowledge Graph can enhance the accuracy of ML classifiers that act as an innovation feature selection model.

Conclusion and Future Work

In this research, we proposed a knowledge-graph model to help machine learning classifiers to predict different Hadith categories. We have seen how the path aggregation technique for recording a convergence among categories and set of words in the knowledge graph with selecting pivot terms led to optimistic results when integrating the proposed model with machine learning classifiers. The experimental results showed that the best result that gave the highest accuracy was random forest RF based on the knowledge-graph model. Overall, the proposed knowledge-graph model improved the performance of all classifiers.

In future work, the general scheme is to expand this work by doubling the volume of data to include most of all the books of the Sunnah of the Prophet. The knowledge-graph technique will also be relied upon and developed into an input to deep learning algorithms with word embedding to improve classification model performance.

ORCID

Mohamed Atef Mosa http://orcid.org/0000-0003-0977-0602

References

Abdelaal, H. M., B. R. Elemary, and H. A. Youness. 2019. Classification of hadith according to its content based on supervised learning algorithms. *IEEE Access* 7:152379–87. doi:10.1109/ACCESS.2019.2948159.

Adeleke, A. O., N. A. Samsudin, A. Mustapha, N. Nawi et al. 2017. Comparative analysis of text classification algorithms for automated labelling of quranic verses. *International Journal on Advanced Science, Engineering and Information Technology* 7(4):1419. doi:10.18517/ijaseit.7.4.2198.

Agrawal, R., T. Imieliński, and A. Swami. 1993. Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD international conference on Management of data, Washington D.C. USA, 207–16.

AL-Kabi, Mohammed N., et al. 2015. Extended topical classification of hadith Arabic text. Int. J. Islam. Appl. Comput. Sci. Technol 3(3): 13-23.

Bahassine, S., A. Madani, M. Al-Sarem, M. Kissi et al. 2020. Feature selection using an improved Chi-square for Arabic text classification. *Journal of King Saud University-Computer and Information Sciences* 32(2):225–31. doi:10.1016/j.jksuci.2018.05.010.

Bakar, M. Y. A., et al. 2018. Multi-label topic classification of hadith of Bukhari (Indonesian language translation) using information gain and backpropagation neural network. In: 2018 International Conference on Asian Language Processing (IALP), IEEE, Telkom University campus, Bandung, Indonesia, 344–50.

Dong, F., and Y. Zhang 2016. Automatic features for essay scoring–an empirical study. In: Proceedings of the 2016 conference on empirical methods in natural language processing, Austin, Texas, USA, 1072–77.

El-Fishawy, Nawal, et al. 2014. Arabic summarization in twitter social network. *Ain Shams Engineering Journal* 5(2): 411–420.

Elghazel, H., A. Aussem, O. Gharroudi, W. Saadaoui, et al. 2016. Ensemble multi-label text categorization based on rotation forest and latent semantic indexing. *Expert Systems with Applications* 57:1–11. doi:10.1016/j.eswa.2016.03.041.

El-Khair, I. A. 2006. Effects of stop words elimination for Arabic information retrieval: A comparative study. *International Journal of Computing & Information Sciences* 4 (3):119–33.

Flachsbart, B., et al. 1994. Using the ID3 symbolic classification algorithm to reduce data density. In: Proceedings of the 1994 ACM symposium on Applied computing, Phoenix Arizona USA, 292–96.

Ghazizadeh, M., et al. 2008. Fuzzy expert system in determining Hadith 1 validity. In *advances in computer and information sciences and engineering*, Dordrecht: Springer, Azadi Square, Iran, 354–59.

Hajjar, M., et al. 2010. A system for evaluation of Arabic root extraction methods. In: 2010 Fifth International Conference on Internet and Web Applications and Services, IEEE, Barcelona, Spain, 506–12.

Harrag, F., and E. El-Qawasmah. 2009. Neural network for Arabic text classification. In: 2009 Second International Conference on the Applications of Digital Information and Web Technologies. IEEE, London, UK, 778–83.

Harrag, F., E. El-Qawasmeh, and P. Pichappan. 2009. Improving Arabic text categorization using decision trees. In: 2009 First International Conference on Networked Digital Technologies. IEEE, Ostrava, Czech Republic, 110–15.

Hassanat, A. B., et al. 2014, August. Solving the problem of the K parameter in the KNN classifier using an ensemble learning approach. *arXiv preprint arXiv*: 1409.0919

Kamsin, A., et al. 2014. Developing the novel Quran and Hadith authentication system. In: The 5th International Conference on Information and Communication Technology for The Muslim World (ICT4M), Kuching, Malaysia, IEEE, 1–5.

Maazouzi, F., and H. Bahi. 2012. Using multi decision tree technique to improving decision tree classifier. *International Journal of Business Intelligence and Data Mining* 7 (4):274–87. doi:10.1504/IJBIDM.2012.051712.

Maraoui, H., K. Haddar, and L. Romary. 2018. Segmentation tool for hadith corpus to generate TEI encoding. In *International Conference on Advanced Intelligent Systems and Informatics*, 252–60. Cham: Springer.

Mosa, M. A. 2019a. Real-time data text mining based on Gravitational Search Algorithm. *Expert Systems with Applications* 137:117–29. doi:10.1016/j.eswa.2019.06.065.

Mosa, M. A. 2020a. A novel hybrid particle swarm optimization and gravitational search algorithm for multi-objective optimization of text mining. *Applied Soft Computing* 90:106189. doi:10.1016/j.asoc.2020a.106189.

Mosa, M. A. 2020b. Data text mining based on Swarm Intelligence Techniques: Review of text summarization systems. In A. Fiori (Eds.), *Trends and Applications of Text Summarization Techniques* (pp. 88-124). IGI Global. http://doi:10.4018/978-1-5225-9373-7.ch004

Mosa, M. A. 2017c June 5. How can ants extract the essence contents satellite of social networks. LAP LAMBERT Academic Publishing, 333032645X.

Mosa, M. A., A. Hamouda, and M. Marei. 2017a. Ant colony heuristic for user-contributed comments summarization. *Knowledge-Based Systems* 118:105–14. doi:10.1016/j.knosys.2016.11.009.

Mosa, M. A., A. Hamouda, and M. Marei. 2017b. Graph coloring and ACO based summarization for social networks. *Expert Systems with Applications* 74:115–26. doi:10.1016/j.eswa.2017.01.010.

Mosa, M. A., A. S. Anwar, and A. Hamouda. 2019b. A survey of multiple types of text summarization with their satellite contents based on swarm intelligence optimization algorithms. *Knowledge-Based Systems* 163:518–32. doi:10.1016/j.knosys.2018.09.008.

Rostam, N. A. P., and N. H. A. H. Malim. 2019. Text categorisation in Quran and Hadith: Overcoming the interrelation challenges using machine learning and term weighting. *Journal of King Saud University-Computer and Information Sciences* 33: 658-667.

Saloot, M. A., N. Idris, R. Mahmud, S. Ja'afar, D. Thorleuchter, A. Gani et al. 2016. Hadith data mining and classification: A comparative analysis. *Artificial Intelligence Review* 46 (1):113–28. doi:10.1007/s10462-016-9458-x.

Suryana, N., F. S. Utomo, and A. Mohd Sanusi. 2018. quran ontology: Review on recent development and open research issues. *Journal of Theoretical & Applied Information Technology* 96: 3.

Zainol, Z., et al. 2016. Discovering "interesting" keyword patterns in Hadith chapter documents. In: 2016 International Conference on Information and Communication Technology (ICICTM), IEEE, 104–08, Vienna, Austria.