IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 21, 2019, accepted September 22, 2019, date of publication September 26, 2019,
date of current version November 11, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2944118

Knowledge Discovery in the Hadith According to the Reliability and Memory of the Reporters Using Machine Learning Techniques

HAMMAM M. ABDELAAL[1], ABDELMOTY M. AHMED[2], WADE GHRIBI[2], AND HASSAN A. YOUNESS ALANSARY[1]
[1]Department of Computers and Systems Engineering, Faculty of Engineering, Minia University, Minia 61519, Egypt
[2]Department of Computers and Systems Engineering, King Khalid University, Abha 62529, Saudi Arabia

Corresponding author: Hammam M. Abdelaal (hammammohamed51@gmail.com)

ABSTRACT Muslims suffer from not knowing the validity of the Hadiths or verification of its degree, which represented the second source of legislation in Islam after the Quran. Although many sites allow users in general and Muslims, in particular, the possibility of verifying the authenticity of the hadith, but it is through an information system which is connected to the database of the Hadith. So, there is no intelligent system that can distinguish the hadith automatically, therefore, in this study, we propose a model that can recognize and categorize the hadith automatically and conclusion the essential features through Hadith classification into Sahih, Hasan, Da'if, and Maudu, based on machine learning techniques. This study is primarily concerned with classifying the hadith according to the memory and reliability of the Hadith's narrators. This classification does not depend only on the text of the hadith, as in the rest of the other Arab documents, but depend on also the Sanad of Hadith. Therefore, this study was conducted on three methodologies; these methodologies help us to obtain an accuracy that is more reliable for hadith classification compared to previous researches in this area. In addition to building a model using the Decision tree technique based on the Sanad of hadith for helping us deciding to judge the validity of the hadith, the accuracy of this classifier reached up to 92.59%. Several Learning algorithms have been used in this study, but we reported the best three classifiers (LinearSVC, SGDClassifier, and Logistic regression), which achieved higher accuracy reached up to 93.69%, 93.51, and 92.27% respectively.

INDEX TERMS Data mining, machine learning algorithms, text preprocessing, feature selection, text classification, prophetic hadith.

I. MOTIVATION AND PROBLEM

There are two motivations for this study Firstly, Texts of the Prophet's Hadiths are fertile grounds for the natural languages processing, knowledge discovery, and the data mining tasks such as the classification [7], [30]. These texts are distinguished by unique linguistic features and the clear link between the word and its meaning. Second: Muslims suffer from not knowing of the validity of the Hadith or verification of its degree, which the second source of legislation in Islam after the Quran. Several sites allow the possibility of verifying the authenticity of the hadith through an information system which connected to the database of Hadith but there is no intelligent system that can distinguish and classify the hadith automatically based on machine learning techniques. In addition, the classification of Hadith to Sahih and Da'if does not depend on the text-only, but also depend on the Sanad of hadith (chain of narrators) and other classification. Therefore in this study, we build a classifier model that can classify and differentiate hadith, according to the reliability and memory of the reporters (narrators), and based on the Takhreej al-hadith ways, using machine learning technique, and finding a relationship among these classifications, using some statistical methods.

The associate editor coordinating the review of this manuscript and approving it for publication was Shuai Liu.

II. INTRODUCTION

Text classification is a task of data mining; it aims to assign automatically selected documents into categories from a pre-defined set of classes [2]. This task is usually solved by combining Information Retrieval and Machine Learning

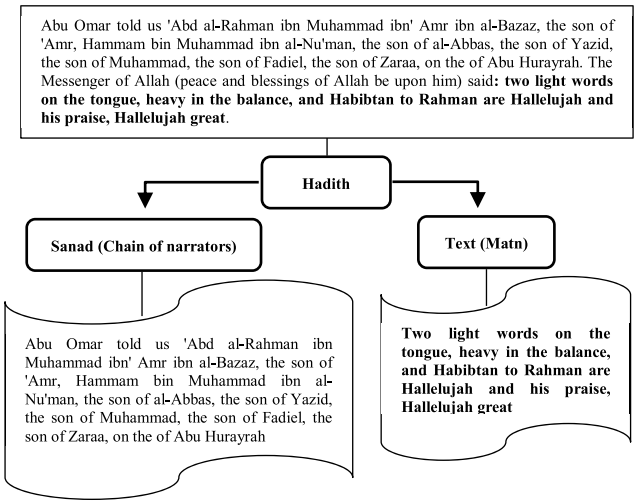


FIGURE 1. Parts of hadith.

techniques. There are two approaches involved in the processing of text classification: The first approach is related to the extract the feature terms which are recognized as effective keywords in the training phase that distinguish each category from the other, and the second approach is concerned with the actual classification of the document using these feature terms in the test phase, that have been used before in training phase. The main goals of the paper are to classify the prophetic hadith into different categories according to the reliability and memory of the reporters, determine performance and efficiency of the categorization model, and try to get a relationship or correlation between hadith classifications.

Hadith consists of two main parts: the text known as the Matn and the Sanad (narrators' chain of hadith) as shown in figure 1. The authenticity of the hadith depends on the reliability of its parts, to judge the degree of hadith, it is necessary to study the Sanad of hadith; therefore this classification has been done based on the text and the Sanad together [1] In this study, the supervised learning algorithms were used to build classification models capable of classifying the hadith into different categories mainly: Sahih, Hasan, Da'if, and Maudu according to the memory and reliability of the narrators. Term Frequency-Inverse Document Frequency (TF-IDF) technique is used to compute the relative frequency for each word in the text of hadith; then a feature selection technique is used to select the most relevant words in each class, after that the dataset of hadith is splitting using cross-validation method. In this method the dataset of hadith is divided randomly into several n blocks, each block of them is held out once to test the classifier and the classifier is trained on the remaining (n-1) to build the classifier [2], [26].

Muslims Scholars give great attention to study the Sanad of Hadith, because it explains to us if Hadith is Sahih or Da'if, so they set some of methods and methodologies like Takhreej al-hadith and the study of al-Asaneed. Takhreej al-hadith is a science interested in studying the Hadiths, and hadith extraction and its authentication based on the study of the case of hadith' narrators. One of the most famous books that investigate in hadith is Sahih Al- Bukhari, Sahih Muslim, Sunan Abu Daoud, Sunan Al Tirmidhi, Sunan Ibn Majah, and Sunan Al Nasai, that represent Sahih and Hasan hadiths; In addition to the science of the wound and modification, which is one of the sciences of the Sunnah Prophet, that identifies the rank of narrators, according to specific terms and concepts. In this study, these criteria have been taken into consideration in addition to some characteristics and conditions that characterize and determine the hadith classification according to the memory and reliability of the narrators. This classification was based on three methodologies.

These methodologies help us to get better accuracy in the classification of hadith. The first methodology is based on the classification of the Hadith based on Takhreej al-hadith ways and the study of narrators, using the algorithm of the decision tree. The second methodology, the hadith has been classified automatically into different categories using supervised learning algorithms, according to what was attributed to the Prophet Muhammad (Saying, Doing, Describing, or Reporting). The third methodology is based on the classification of Hadith for the text and the Sanad together. The Sanad of hadith is converted to feature terms which are composed of narrators' names using N-grams method to configure a sequence of narrators and it's distinguished in patterns. This distinction of narrators' names is important to identify and evaluate the narrators. After these statistical methods were used to find correlations between the classification of Hadith according to the reliability and memory of the reporters and classification of Hadith according to what was attributed to the Prophet, such as the correlation of the Sahih hadith or Da'if with Saying, Doing, Describing and Reporting. The remainder of this paper is organized as follows: Section 2 shows the related works.

The methodology of this study is presented in Section 3. Section 4 shows the proposed system for this study. Section 5 shows the Classification of Hadith according to the Reliability and Memory of the Reporters. Relationship between these classifications is illustrated in Section 6. The Experimental results and Evaluation are shown in Section 7. Finally, the conclusion is presented in section 8.

III. RELATED WORKS

Many type of researches have been studied to solve the problem of Arabic text classification, while few of researches have been applied to knowledge discovery in the hadith, and its classification, which can be used as a source in knowledge discovery; it contains unique linguistic features other than other Arabic texts. In this section, we present some of the previous studies and its results in prophetic hadith classification as shown in Table 1, like the following.

Mohammed *et al.* [3] evaluated the automatic classification of Islam prophet sayings based on several possible categories. Their study aims to classify the hadith into one of five categories (books) mainly: Fasting, Zakat, four algorithms are

TABLE 1. Results of previous works related to Hadith classification.

Approach Published by	Methods	Categories	Dataset	Performance
Khitam Jbara. (2010), [4]	Stem Expansion Classification (SEC), Cosine Coefficient	13 Subjects (topics)	1300 Hadiths from Thirteen Books of Sahih Bukhari	Accuracy: 73%
Bilal et al. (2012), [5]	Cloud-based Expert System	A various topic of Hadith	Example of hadith	Not reported
Aldhaln et al. (2012), [7]	Decision trees Naive Bayes	Sahih, Hasan, Da'if and, Maudu	999 Hadiths from Sahih Al-Bukhari,, Silsilat Al-Ahadith Al-Daeifah Wal Al-Mawdhuah	Accuracy: 97%
Najeeb. (2014), [6]	Term Indexing, Associative Classification	Sahih and Da'if	Not reported	Accuracy: 72%
Kabi et al. (2015), [3]	Bagging, LogiBoost SVM	Five Categories	Collect 793 Hadith Distributed among Five classes	Accuracy:: 59.30%

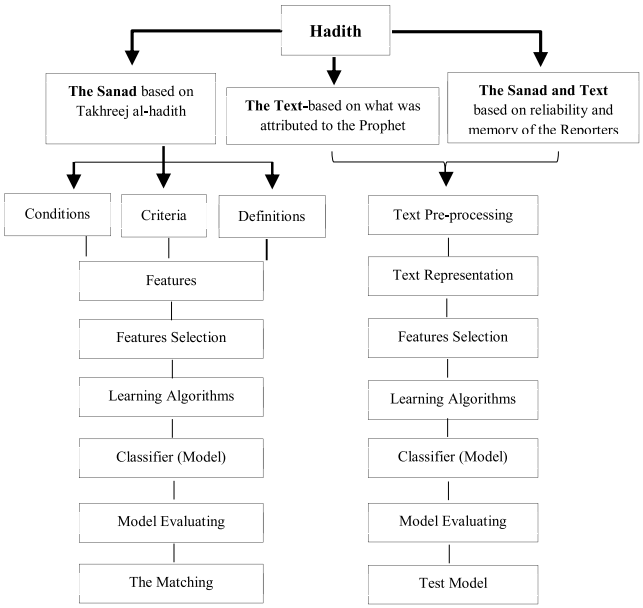


FIGURE 2. Methodology of hadith classifications.

used in this study mainly: Naïve Bayes, Bagging, LogiBoost, and SVM, This study is based on Sahih Bukhari book.

Jbara [4] study is conducted to classify the hadith to one of the predefined classes (books). In this study, a collection of Hadiths containing about 1300 Hadiths from Thirteen Books of Sahih Bukhari are selected and each Hadith is assigned to one chapter (Kitab). The results show that Stem Expansion Classification (SEC) performed better in classifying hadiths against existing classifications methods according to the most reliable measurements mainly: recall, precision, and F-Measure) in text classification field.

Bilal and Mohsen [5] they introduced a Cloud-based Distributed Expert System to classify Prophet Mohammed (PBUH) sayings that use Hadith science to classify them among 24 types from seven broad categories. The capabilities of this system are not limited to classification, but it can identify fabricated and authentic sayings. This study presents the relationship and mapping of the expert system technology onto Hadith sciences, and technicalities involved in designing of the Muhadith expert system.

Najeeb [6] proposed a new classification approach that distinguishes between authenticated (Sahih) and weak (Da'if) Hadiths, for judgment of Hadith and differentiate between the accepted and rejected Hadith. He explained in his study; that there is a big opportunity to build an automatic information system to classify Hadith to Sahih or Da'if using Associative Classification technique known as Association rule mining (ARM), which aims to discover the relationship between the features in order to define a set of classification rules. In this study, no explicit accuracy rate has been reported.

Aldhlan *et al.* [7] study the Hadith Classifier that built using the Decision Tree algorithm. A novel mechanism called missing data detector (MDD) was employed to handle these missing data. This mechanism simulated the Isnad verification methods in Hadith science. The results of the research were compared with the sourcebooks, concurrently with the point of view of the experts in Hadith science. The attributes of the instances originally were obtained from the source books. Whilst some attributes were indicated as null values or missing values. The findings of the research showed that the performance of the decision tree had a significant effect on missing data detector.

Saloot *et al.* [8] presents all previous studies on Hadith classification, and data mining techniques, whether the hadith was classified according to Sahih (authentic) and Da'if (unauthentic) or according to its content (topic), indicating the methods used, the data set, the categories and the results of each study.

IV. STUDY METHODOLOGY

This study is adopted on three methodologies to classify the Hadith according to the reliability and memory of the reporters (narrators) as shown in Figure 2. The first methodology classification the Hadith based on Takhreej al-hadith ways and the case study of narrators, using the algorithm of the decision tree. The decision tree was chosen for its easy building in addition to its ability to identify the most important features in the dataset and the effective features which the class of hadith is depend [9]. In the second methodology, the hadith has been classified into different categories using supervised learning algorithms, according to what was attributed to the Prophet (Saying, Doing, Describing, or Reporting), all these categories fall under authenticated hadiths, and they are more generalized than classification of hadith according to the reliability and memory of the reporters (narrators). This methodology consists of three main stages: firstly, pre-processing stage, secondly, text modeling stage, and finally text classification. The pre-processing stage includes: tokenization, stop word removal, normalization, and stemming. The text modeling stage includes: Text

representation known as vector space model, to calculate the term weighting for each word and feature selection to choose the most relevant words that distinguish between classes in the dataset. The text classification stage includes a model (classifier) building and model (classifier) evaluation and testing using unseen examples of hadith.

In the third methodology, the Hadith is classified based on the text and the Sanad together, Noting that, combination the Sanad of Hadith (hadith' narrators) and the text gives higher accuracy in the classification. In actually this matches what hadith scholars said that judge the degree of the hadith depends on the Sanad and text. This methodology consists of the same previous stages as in the second methodology, in additional the N-grams technique was used to configure a sequence of narrators and it is distinguished in patterns. This distinction of narrators' names is important to identify and evaluate the narrators; each pattern of them indicates a specific class of hadith classes and trains the learning algorithms. After these statistical methods were used to find correlations between the classification of Hadith according to the reliability and memory of the reporters and classification of Hadith according to what was attributed to the Prophet, such as the correlation of the Sahih hadith or Da'if with Saying, Doing, Describing and Reporting. Table 1 shows a sample of prophetic hadith.

V. PROPOSED SYSTEM

The proposed system consists of four stages; the first one is the data pre-processing, followed by the learning algorithms stage, the input of this stage is a set of pre-classified hadith that related to the classification of hadith according to what was attributed to the Prophet, the output is the classifier model. The third stage is the classifier evaluation, and finally classifier testing, using unseen examples of hadith that related to classification of hadith according to the Reliability and Memory of the Reporters, from the outputs of this classifier, we can find out a relationship between these classifications as shown in Figure 3, which shows the Outline of Hadith Classification Process.

VI. CLASSIFICATION OF HADITH ACCORDING TO THE RELIABILITY AND MEMORY OF THE REPORTERS

The classification of the Hadith into Sahih, Hasan, Da'if, and Maudu, according to the Reliability and Memory of the hadith' narrators is not necessarily easy, but requires deep knowledge in the hadith sciences, because this classification does not depend only on the text of the hadith as like other Arabic texts, but also on the Sanad [10], [11] Therefore, in this study, we adopted three methodologies to classify the Hadith, as the previous that mentioned before in section 3.

A. THE SANAD BASED ON TAKHREEJ AL-HADITH

The classification of the Hadith depends mainly on the Sanad; it is the main part in hadith and it interested in studying the status of narrators. In this method, the Hadith was classified based on the Sanad, according to a set of conditions and

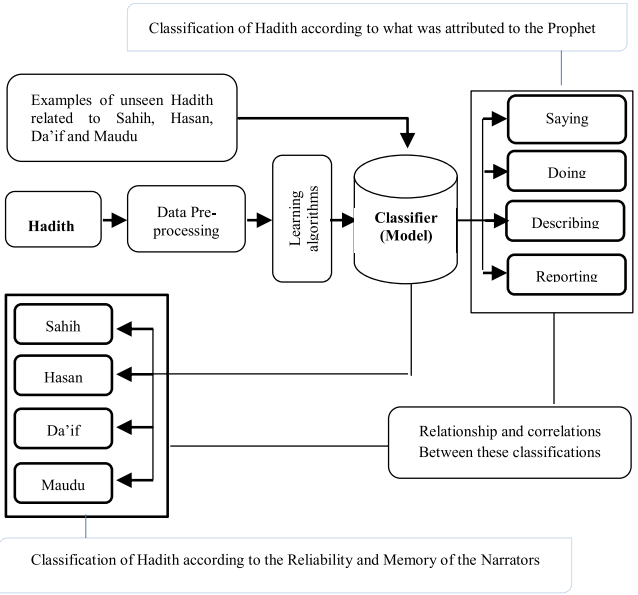


FIGURE 3. Outline the hadith classification process.

TABLE 2. Main features of a training dataset of hadith.

ID	R	AN	N	SC	SBA	JN	Class
1	Bukhari\ Muslim	complete	confidence	strong	yes	yes	Sahih
2	Other	complete	confidence	strong	yes	yes	Sahih
3	Other	complete	Sadouk	strong	yes	yes	Sahih
4	Other	light	Sadouk	strong	yes	yes	Hasan
5	Other	light	Unknown	strong	yes	yes	Hasan
6	Other	unknown	confidence	strong	yes	yes	Sahih
7	Other	unknown	Sadouk	strong	yes	yes	Hasan
8	Other	unknown	Unknown	strong	yes	yes	Da'if
9	Other	unknown	Unknown	weak	yes	yes	Da'if
10	Other	unknown	Unknown	weak	no	no	Maudu
11	Other	unknown	Liar	weak	no	no	Maudu
-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-

criteria which through the validity of the Hadith is judged, and based on the definition of each class like Sahih, Da'if [12], [13] One of the most important of these conditions and criteria is the study of the case of narrators such as justice, honesty, trust, the degree of conservation, adjusts of the narrator, etc.

These conditions and criteria were formulated to a set of features to train the learning algorithm for recognizing the class of hadith; each feature of them has a set of possible values which represent the dataset of hadith as shown in Table 2. Also every hadith - excepting the hadiths of al-Bukhari and Muslim is not judged until after a study of its Isnad and its Matn according to the rules that have put by hadith scholars [14], [15], these conditions have been represented as shown in Table 2, as the following: the Sanad connection (SC), the justice of the narrators (JN), adjust the narrators (AN), safety of anomalies, and safety of the bug (SBA), addition to narrator of hadith (N), it is the first narrator for hadith,

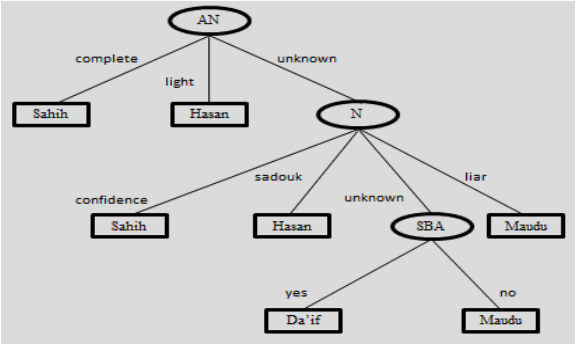


FIGURE 4. Decision tree for classification the hadith according to reliability and memory of the reporters.

TABLE 3. Features ranking according to information gain technique.

Attribute	AN	N	JN	SC	SBA	R
Ranking	0.8233	0.7106	0.6052	0.5844	0.3831	0.061

and the Reference (director) of hadith (R), the most famous directors of the Hadith are Imam al-Bukhari, Imam Muslim, Abu-Dawood, Ibn-Majah, Al-Tirmidhi [16]. The information gain technique is used to determine the feature that has the highest gain and the most important feature that are relevant to target data and ignore the data that is irrelevant to the target data. The highest features are ranking as the follows: AN, N, JN, SC, SBA, and R as shown in Table 3.

The decision tree has been generated as shown in Figure 4 using ID3 (Iterative Dichotomiser) [28] based on information gain technique (IG) to find the best split the tree and to identify the effective feature that helps taking the decision to judge the validity of the hadith according to equation1, equation2, and equation3. IG is used to measure the dependence between features and labels and calculates the gain between the (i-th) feature f_i and the class labels. After calculating both of the expected information needed to classify the hadith in dataset D is given by equation 1, and the expected information required for each feature according to equation 2.

The gain of each feature is calculated according to equation 3. The decision tree has been evaluated and matched with these rules by accuracy is 92.59%. The decision tree has been chosen because it is very suitable for the classification process that able to identify the most important features in the training dataset and it can deal with missing values; in addition to its easy construction that provides a rapid and effective method of classifying the hadith. Finally, the tree is tested by matching it with the conditions and rules of hadith Degrees.

$$Info(D) = -\sum_{i=1}^{n} P_i \, log_2 \, (P_i) \tag{1}$$

where Info(D) also known as the entropy of D, it is the average amount of information needed to determine the class of hadith in D (Dataset), P_i is the probability that an arbitrary hadith in D belongs to a class of hadith, and i the total number of categories [2], [17].

$$Info_f(D) = \sum_{j=1}^{n} \frac{|Dj|}{|D|} \, x \, Info \, (Dj) \tag{2}$$

$Info_f$ (D) is the expected information required for each feature to classify a hadith from D based on the partitioning by $\frac{|Dj|}{|D|}$ that acts as the weight of the jth partition.

$$Gain(f) = Info \, (D) - Info_f(D) \tag{3}$$

Note that, the construction of the tree is consistent with the classifications of hadith. The most important that distinguishes Sahih hadith about the Hasan is adjusting the narrators. Hasan hadith contains all characteristics of Sahih but with less adjust. Da'if hadith is one who misses one or two conditions of Sahih or Hasan. Maudu is outside the scope of hadith classification because it is not the Say of the Prophet Muhammad, there is one or more of its narrators are accused of lying.

Algorithm 1 Decisions

```
From this tree, we can conclude these decisions (rules)
if (the hadith found in al_bukhari or Muslim Sahih) then
    the hadith is Sahih
else
    check the status of Adjust_Narrators.
if (Adjust_Narrators = complete) then
    the hadith is Sahih
else if (Adjust_Narrators = light) then
    the hadith is Hasan
else
    check the status of the Narrator.
if (Narrator = confidence) then
    the hadith is Sahih
else if (Narrator = Sadouk) then
    the hadith is Hasan
else if (Narrator = liar) then
    the hadith is Maudu
else
    check the status of Safety_bug_abnormality.
if (Safety_bug_abnormality = yes) then
    the hadith is Da'if
else
    the hadith is Maudu.
```

B. THE TEXT-BASED ON WHAT WAS ATTRIBUTED TO THE PROPHET

In this methodology the hadith has been classified into different categories mainly: Describing, Saying, Reporting and Doing; according to what was attributed to the Prophet Muhammad, based on the text only. Table 4 shows examples of hadiths that are related to each category, given the classification that depends on the text of hadith only; the hadith has been classified like other Arabic texts rest. Therefore, this

TABLE 4. Examples of hadith that are related to each category.

English Hadith	Arabic Hadith	Class
Allah's Apostle prayed the `Isha' prayer and pray eight rak`at and then pray two rak`at while sitting. He then prayed two rak`at in between the Adhan and Iqama and he never missed them	صلى النبي العشاء ثم صلى ثماني ركعات وركعتين جالسا وركعتين بين النداءين ولم يكن يدعهما أبدا	Reporting
He performed ablution thus: He washed his face thrice, and his hands twice, then passed his wet hands lightly over the head from front to rear and brought them to the front again and washed his feet.	توضأ فغسل وجهه ثلاثا ويديه مرتين مرتين ومسح برأسه فأقبل به وأدبر وغسل رجليه	Doing
Beware of narrating too many hadiths from me. Whoever attributes something to me let him speak the truth faithfully. Whoever attributes to say something that I did not say let him take his place in Hell	إياكم وكثرة الحديث من حدث عني فلا يقول إلا حقا من قال علي ما لم أقل فليتبوأ مقعده من النار	Saying
Allah's Apostle was the most handsome of all the people and had the best appearance. He was neither very tall nor short.	أحسن الناس وجها أحسنه خلقا ليس بالطويل البائن ولا بالقصير	Describing

TABLE 5. Number of collected hadith per category.

Classification	Class A				Class B			
Class	Describing	Doing	Saying	Reporting	Da'if	Hassan	Maudu	Sahih
No. of Hadith	1759	1732	1741	1625	1294	1394	1038	1384
Total	6857				5110			

text has been processed to convert it to a form that is suitable for classification tasks.

Hadith Classification Framework consists of four main stages: the hadith collection stage, text pre-processing stage, the text modeling stage, and text classification. The pre-processing stage includes remove the Sanad of hadith, tokenization, stop word removal, normalization, and stemming. The text modeling stage includes vector space model construction or term weighting, features selection, and the hadith classification stage covers classification model building, classification model evaluation, and classification model testing.

1) HADITH COLLECTION

In this study a large number of prophetic hadith was collected from different sources, that collected from famous books in hadith such as Sahih Imam al-Bukhari, Muslim, Tirmidhi and al-Nasaa'i, Ibn Majah, Abu Dawood, and a Silsilat al_hadith al Da'if and Maudu by imam al-Albani, in addition to the encyclopedia of hadith al-Shareef, https://islamweb.net/hadith/hadithsrch.php, in order to study the classification of Hadith according to what was attributed to the Prophet and according to the reliability and memory of the reporter. The number of hadith used in this study as shown in Table 5, where Class A refers to the number of hadiths that are related to the classification of hadith according to what was attributed to the Prophet, and Class B refers to the number of hadiths that are related to the classification of hadith according to the reliability and memory of the reporters. After collected these Hadiths we come to pre-processing technique.

2) HADITH TEXT PRE-PROCESSING

Pre-processing is one of the main steps in a text classification process, aims to converts the text to a form that is suitable for the classification process. These pre-processing include linguistic tools such as normalization, tokenization, stop word removal and stemming [19], [20], to reduce the ambiguity of words and increase the accuracy and effectiveness of the classification. Previous experimental results revealed that pre-processing techniques have a significant impact on the Arabic texts classification accuracy [18], [19].

a: REMOVE THE SANAD

The Sanad (chain of narrators) of hadith is removed from the hadith such as 'إبراهيم حدثنا الحميدي عبد الله بن الزبير قال الخطاب رضى حدثنا سفيان قال حدثنا يحيى بن سعيد الأنصاري قال أخبرني محمد بن سمعت عمر بن وقاص الليثي يقول سمع علقمة بن أنه المنبر التيمي على تعالى عنه الله', because this classification does not depend on it, and it has no effect on the accuracy of the classification. Unlike the classification of Hadith according to the Reliability and Memory of the Reporters, it represents the main part of this classification. Also, the special characters, diacritics and punctuation marks are removed because they are not useful in the classification process.

b: NORMALIZATION

The word normalization aims to normalize certain letters that have different forms in the same word to one form, such as Replace (ﺃ, ﺁ, ﺇ) with (ﺍ), (ﻯ) with (ﻱ), and (ﺓ) with (ﻩ) [19].

c: TOKENIZATION

Tokenization aims to divide the text into tokens. Words are often separated from each other by blanks such as white space, semicolons, and commas.

d: STOP WORD REMOVAL

Stop word removal aims to eliminate of insignificant words, such as so which appear in the sentences frequently and do not have any meaning or impact about the content, such as, in, on, to, etc.[19].

3) STEMMING

The stemming technique aims to remove all affixes such as prefixes, infixes, and suffixes from the word, to reduce the appearance of different forms for the word that reflects the same meaning, such as learns, learning, learned, and learn us, all these words share the same abstract meaning. Stemming makes the processes less dependent on particular forms of words and reduces the potential size of features, which, in turn, improve the accuracy of the classifier in the text classification [20], [21].

TABLE 6. Format file includes Boolean approach.

F1	F2	F3	- - -	Fn	Class
1	0	1	- - -	1	Reporting
0	0	1	- - -	0	Saying
0	1	0	- - -	0	Doing
1	1	1	- - -	1	Describing
0	0	0	- - -	0	Saying

C. TEXT MODELING

This stage includes two techniques, the first technique is the vector space model known as Term Weighting and the second technique for features selection.

1) VECTOR SPACE MODEL

Vector space model (VSM) known as term weighting that is an algebraic model for representing text documents as a vector or for converting full text into vectors such as index terms, which frequently used in information retrieval, indexing and text classification [22], [23] Each dimension matches to a separate word, where if a word occurs in the text, then its weighting value in the vector is non-zero.

Assume we have a collection of documents D containing N documents, such as $D = \{d_0, d_1 \ldots dn_{-1}\}$, each document of them d_i that contains a collection of terms t will be represented as vectors in VSM as follows:

$d_{ij} = (t_{1j}, t_{2j}, \ldots\ldots, t_{ij})$, $j = 1, 2, \ldots\ldots$, m, where m refers to the number of distinct words in the document di.

There are different methods used to calculate the weight for each word in the text such as

- Boolean algebra indicates the absence (0) or the presence (1) of a word in the dictionary of words that contained from the dataset [24] It is a method of representing a word using only two values (1 or 0) in matrix form as shown in Table 6.
- Term Frequency (TF), is a number of times that the term appears in a document, to see if you're using this term too much or few in the document, it is calculated according to the following equation [21].

$$TF_{i,j} = \frac{n_{i,j}}{\sum n_{k,j}} = \frac{f_{w,t}}{\max\{f_{w',t} : w' \varepsilon t\}} \tag{4}$$

$n_{i,j}$: is the number of occurrences of the word (w_i) in the text of hadith (t_j).
$\sum n_{k,j}$: is the sum of the number of all words (w_k) in the text of hadith (t_j).
$TF_{i,j} = 1$ if w occurs in t and 0 otherwise.

- Inverse Document Frequency (IDF), is a measure of how much information the term provides, to see if you're using this term too much or few in all documents (common or rare across all documents). It equal to the logarithm of quotient divided the total number of documents by the number of documents containing the term.

$$IDF(w, t) = log\frac{N}{n} = log(\frac{N}{|\{t \varepsilon D : w \varepsilon t\}|}) \tag{5}$$

where, N is the total number of texts in the dataset of hadith N= |D|. $|\{t \varepsilon D : w \varepsilon t\}|$; n is the number of texts where the w appears or contains it, $TF_{i,j} \neq 0$, if the term is not found in the dataset, this will lead to a division-by-zero. It, therefore, adjusts the denominator to become $|\{t \varepsilon D : w \varepsilon t\}| + 1$.

- Term Frequency-Inverse Document Frequency (TF-IDF), it is the product of two statistics TF and IDF; that is used to determine the frequency of each word in a hadith text according to equation 3 after computing the TF and IDF according to equation 4 and equation 5 respectively [8] But the most common method is TF-IDF, is a method statistic aims to reflect how important a word is to text in a dataset, which helps us to determine the important words in a documents collection for classification purposes [25] The value of TD-IDF increases proportionally to the number of times for word appear in the text and it is offset by the number of texts in the corpus that contain the word.

$$TD - IDF = TF_{i,j} * IDF(w, t) = \frac{n_{i,j}}{\sum n_{k,j}} x log \frac{N}{|\{t \varepsilon D : w \varepsilon t\}|} \tag{6}$$

TABLE 7. TF -IDF FOR sample of hadith after preprocessing.

TF-IDF	After Pre-processing	Hadith Text
0.12690695609092 = اذا 0.346076882942 = عطس 0.164667193755 = احد 0.3556429686342 = حمد 0.1815843076780 = الل 0.7489506344853 = شمتو 0.237228807694 = فان 0.1428912509418 = لم 0.19983212770594= فلا	اذا عطس احد حمد الل شمتو فان لم حمد الل فلا شمتو	اذا عطس احدكم فحمد الله فشمتوه فان لم يحمد الله فلا تشمتوه If anyone of you sneezes, let him say 'Al-hamdu Lillaah', and let his brother or his companion say, 'Yarhamuk Allah', And if he says to him, 'Yarhamuk Allah', let him say, 'Yahdeekum Allaahu wa yusliha baalakum
0.257904651993 = كلم 0.341066815221 = حبب 0.142089997310 = الى 0.242657567042 = رحم 0.248590785139 = خفف 0.128977476187 = على 0.312933817740 = لسن 0.270405386487 =ثقل 0.107712215296 = في 0.319738225507 = ميز 0.493055229887 = سبح 0.205710020932 = الل 0.201447260112 = حمد 0.220842965694 = عظم	كلم حبب الى رحم خفف على لسن ثقل في ميز سبح الل حمد سبح الل عظم	كلمتان حبيبتان إلى الرحمن خفيفتان على اللسان ثقيلتان في الميزان سبحان الله وبحمده سبحان الله العظيم Two words Habibtan to Rahman, light them on the tongue and heavy in the balance Hallelujah and praise Hallelujah great

Table 7, shows two examples of hadith related to the saying class, after removing the Sanad from hadith and preprocessing for the text to calculate the term weighting for each word in the text of hadithusing TF-IDF.

TABLE 8. SAMPLES of words that have the highest gain ratio.

Word	Chi-squared	Gain Ratio	Information Gain
توضأ	1359.703	0.37604	0.13096
ثلاثا	1318.816	0.28913	0.12219
وجهه	954.0565	0.24430	0.09295
رأسه	837.1903	0.18982	0.07550
غسل	738.0902	0.28623	0.07082
عشرة	552.9674	0.21699	0.05107
النار	487.2345	0.23486	0.04741
ركعة	475.5162	0.20887	0.04276
يده	463.9997	0.17811	0.08614
مرة	412.833	0.25304	0.03883
سجد	375.3573	0.24115	0.03642
اليسرى	371.4237	0.25005	0.03513
الليل	323.3901	0.21391	0.03131
الوضوء	285.2804	0.20767	0.02676
الصلاة	275.8386	0.16525	0.02869
السماء	268.7544	0.15941	0.02817

2) FEATURE SELECTION

After Pre-processing, and text representation using the vector space model to convert the text of hadith to features, the features selection technique is used. The feature selection (FS) is an effective method to solve high dimensional data problem, by removing undesirable and redundant data to improve the classification accuracy which helps us to select the important features that are relevant to target data, and ignore the data that is irrelevant to the target data [31], in additional to; enable the learning algorithm to train faster [23]

Several different methods are used to select the best features such as Information Gain, Latent semantic analysis, Gain Ratio, Symmetrical uncertainty, probability ratio, odds ratio, Fisher score, Chi-Square, Gini Index, Principal Component Analysis, etc. Taking into account the following constraints in the selection of feature 'Vectorizer = TfidfVectorizer (min_df = 10, max_df = 0.75, ngram_range = (1, 3))', means that ignore terms that appear in less than 10 of the hadith texts, ignore terms that appear in more than 75% of the hadith texts and the lower and upper confines of the range of n-values for different n-grams to be extracted. All the values of parameter n such that $1 <= n <= 3$ will be used.

Information Gain, Chi-Square and Gain Ratio techniques are used as feature selection to identify the words that have highest ranking in hadith dataset, they statistic measures to determine the important words in the dataset as shown in Table 8 that contains some words that have highest ranking for each technique. The gain ratio is defined from equation 9, after computing the gain of each attribute A according to equation 8, the attribute with the maximum gain ratio is selected as the splitting attribute. Also, Chi-Square is defined from equation 7, and finally, Information Gain is defined according to equation 1.

$$X_C^2 = \sum \frac{(Observed\ values - Expected\ values)\,2}{Expected\ values} = \sum \frac{(O_i - E_i)2}{E_i} \tag{7}$$

X_C^2 Is the square of the difference between the observed (O) and expected (E) values and divide it by the expected value into one or more classes (C).

The gain ratio is a modification of the information gain that reduces its bias on high-branch attributes that computed according to equation 8. The attribute with the highest gain ratio is selected as the best attribute for splitting according to equation 9.

$$SplitInfo\ A(D) = -\sum_{j=1}^{n} \frac{|Dj|}{|D|} x log_2(\frac{|Dj|}{|D|} \tag{8}$$

$$Gain\ Ratio(A) = \frac{Gain(A)}{SplitInfo\ (A)} \tag{9}$$

D. TEXT CLASSIFICATION

Text classification is a process of assigning a text document to one or more predefined classes based on their content, it is a technique of predictive data mining techniques [2], based on supervised learning algorithms, and includes three steps mainly: Building, Evaluating and classifier Testing [29].

1) CLASSIFIER BUILDING

Building the classifier aims to train the algorithm (classifier) to recognize patterns, main features for each category and makes predictions from the training dataset, to predict the category of hadith. The dataset of hadith is 6857 sample, divided into 10 folds, each fold of them (685) is held out once to test the classifier, and the classifier is trained on the remaining (6170) to build the classifier. Different algorithms have been used in this study, to build a model that can classify the category of hadith.

2) CLASSIFIER EVALUATION

Classifier accuracy is evaluated using many measures mainly, Recall, Precision, and F1-score, after compute the true positives rate (TP), true negatives rate (TN), false negative rate (FN) and false positive rate (FP) [32] TP refers to the number of hadiths which are correctly assigned to a given category. TN refers to the number of hadiths which are not correctly assigned to a given category. FP refers to the number of hadiths which are falsely assigned to the category, and FN refers to the number of hadiths which are not falsely assigned to the category [17], [27]. TP and TN are correctly classifications; FP and FN are incorrectly predicted class as shown in Figure 5.

TP rate is TP divided by the total number of positives

$$(TP + FN) = \frac{TP}{TP + FN} \tag{10}$$

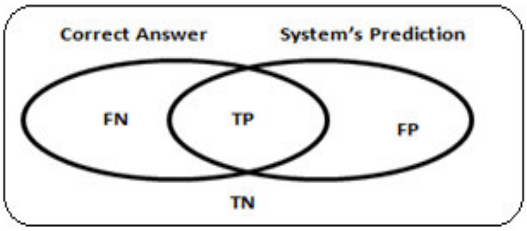


FIGURE 5. Rate of correctly predicted classes.

FP rate is FP divided by the total number of negatives

$$(\mathrm{FP}+\mathrm{TN}) = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}. \tag{11}$$

We can compute the precision, recall, and F1-score from equations 12, 13, and 14 respectively.

Overall accuracy is computed by the number of correct classifications divided by a total number of classifications according to equation 15.

To know how to evaluate the performance of a classifier and determine the number of properly classified and incorrectly classified Hadiths necessity understand the confusion matrix that will be discussed in section 7.2

$$\begin{aligned}\text{Precision} &= \frac{\text{Relevant} \bigcap \text{Retrieved}}{\text{Retrieved}} \\ &= \frac{\text{No. of system's correctpredictions}}{\text{No. of systemsoutputs}} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}\end{aligned} \tag{12}$$

$$\begin{aligned}\text{Recall} &= \frac{\text{Relevant} \bigcap \text{Retrieved}}{\text{Relevant}} \\ &= \frac{\text{No. ofsystem's correct predictions}}{\text{No. of correct answers}} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}.\end{aligned} \tag{13}$$

$$\text{F1-score} = \text{F-Measure} = \frac{2(\text{Precision} * \text{Recall})}{\text{Precision} + \text{Recall}} \tag{14}$$

$$\text{Accuracy} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{15}$$

3) CLASSIFIER TESTING

After we train the classifier and its evaluating, we used the trained classifier to predict examples of hadith that are unseen data (unlabeled), to test, and measure the accuracy of the classifier in the hadiths classification that is untrained on them.

E. THE SANAD AND TEXT-BASED ON RELIABILITY AND MEMORY OF THE REPORTERS

In this methodology, the hadith has been classified based on the Sanad (chain of narrators) and the text together according to the reliability and memory of the Reporters as shown in Table 9. Text of hadith has been Pre-processing by using some methods that are used in the second methodology, but the Sanad was not removed as in the second methodology. Sanad of hadith is taken as the main feature to distinguish between hadith classes, to contain it the narrators' names; therefore distinction of narrators' names is important as identifying and evaluating the narrators which are a very important part in Hadith classification according to Reliability and Memory of the Reporters. The N-grams technique is used to extract the features from the chain of narrators, to appear as a pattern to learn the classifier.

TABLE 9. Sample of the hadith contains the text, Sanad and its narrator.

Hadith	The Sanad	Text	Narrator
حدثنا الحميدي عبد الله بن الزبير قال حدثنا سفيان قال حدثنا يحيى بن سعيد الأنصاري قال أخبرني محمد بن إبراهيم التيمي أنه سمع علقمة بن وقاص الليثي يقول سمعت عمر بن الخطاب رضى الله تعالى عنه على المنبر قال سمعت رسول الله صلى الله عليه وسلم يقول إنما الأعمال بالنيات وإنما لكل امرئ ما نوى فمن كانت هجرته إلى دنيا يصيبها أو إلى امرأة ينكحها فهجرته إلى ما هاجر إليه.	حدثنا الحميدي عبد الله بن الزبير قال حدثنا سفيان قال حدثنا يحيى بن سعيد الأنصاري قال أخبرني محمد بن إبراهيم التيمي أنه سمع علقمة بن وقاص الليثي يقول سمعت عمر بن الخطاب رضى الله تعالى عنه.	إنما الأعمال بالنيات وإنما لكل امرئ ما نوى فمن كانت هجرته إلى دنيا يصيبها أو إلى امرأة ينكحها فهجرته إلى ما هاجر إليه	عمر بن الخطاب

TABLE 10. Sample of the Sanad after applying tri-grams.

The Sanad	Narrators	Tri- Grams
حدثنا الحميدي عبد الله بن الزبير قال حدثنا سفيان قال حدثنا يحيى بن سعيد الأنصاري قال أخبرني محمد بن إبراهيم التيمي أنه سمع علقمة بن وقاص الليثي يقول سمعت عمر بن الخطاب رضى الله تعالى عنه على المنبر	الحميدي عبد الله بن الزبير – سفيان- يحيى بن سعيد الأنصاري - محمد بن إبراهيم التيمي - علقمة بن وقاص الليثي - عمر بن الخطاب	الحميدي عبد الله الزبير، عبد الله الزبير سفيان، الزبير سفيان يحيى، سفيان يحيى سعيد، يحيى سعيد الأنصاري، سعيد الأنصاري محمد، الأنصاري محمد إبراهيم، محمد إبراهيم التيمي، إبراهيم التيمي علقمة، التيمي علقمة وقاص، علقمة وقاص الليثي، وقاص الليثي عمر، الليثي عمر الخطاب.

The Sanad of hadith is converted to feature terms which are composed of narrators' names using N-grams method. There are three techniques of N-grams on word-level mainly: Uni-gram, bi-gram, and tri-gram. Tri-gram is applied, which are commonly used in text classification, and it is effective as a language-independent method because it not depends on the meaning of the language. In tri-gram, every three words in the Sanad is used as a feature, as shown in Table 10, where each feature term, it corresponds to three names of narrators in a specific sequence. Then each feature is weighted using TD-IDF method to know the feature profile for each category in the training stage.

In the test stage, to classify unseen examples of hadith will be pre-processed and represented by using the same pre-processing and representation methods that are used in training stage to convert the text of hadith and the Sanad to a set of features, then applying feature selection methods to use these features that have been selected in matching classification with the features of training.

VII. RELATIONSHIP BETWEEN CLASSIFICATIONS

In this study, we try to discover a relationship between the classifications of hadith according to what was attributed to the Prophet and according to the Reliability and Memory of the Reporters using two approaches. The first is through the outputs of the classification model. The second is through recognizing speech parts and identifying it as a verb, adverb, noun, adjective, etc., using Part of Speech (POS) Tagging technique, based on some statistical methods.

TABLE 11. Numbers of hadiths that belong to each class.

Class	Describing	Doing	Reporting	Saying	Grand Total
Da'if	9	3	22	66	100
Hassan	5	14	13	68	100
Maudu	12	3	31	54	100
Sahih	9	2	22	67	100
Grand Total	35	22	88	255	400

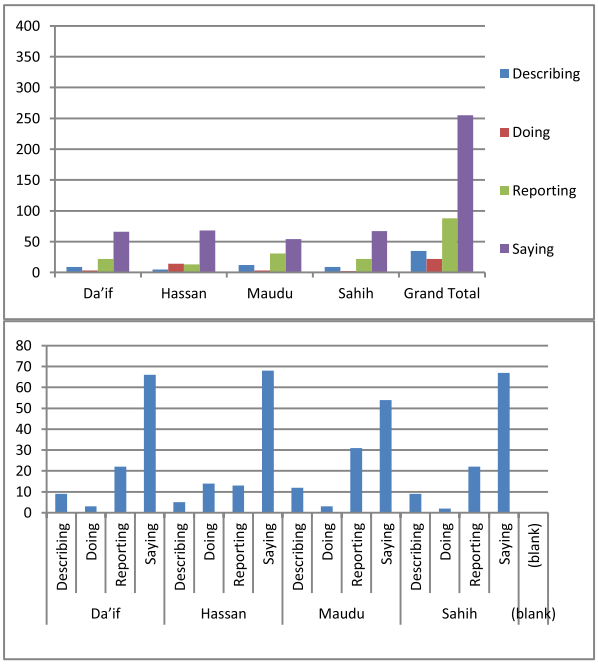


FIGURE 6. Number of hadiths in each class.

A. THE OUTPUT OF THE CLASSIFICATION MODEL

After building these classifiers using supervised learning algorithms according to what was attributed to the Prophet, the best algorithm of them has applied to test 400 sample of hadith (unseen sample) divided to 100 sample per category from dataset of hadith that belong to the classification of hadith according to the reliability and memory of the reporters, to find the degree of interdependence and overlap of these classifications. Then the output of this classifier (algorithm) is tested using the PivotTable function that is one of Excel's most powerful features and allows you to extract the significance from a large data set as shown in Table 11 and Figure 6.

From Table 11, we noting that a large number of hadiths according to the reliability and memory of the reporters found in Saying class, followed by Reporting, then Doing, and finally the Describing class. Then the large number of hadiths that belong to saying class found in Hasan, followed by Sahih, then Da'if, and finally, the Maudu class as shown in Figure 6.

B. PART OF SPEECH (POS-TAGGING)

POS Tagging is applied to the text of hadith to obtain a relation between the class of hadith and its parts (verb, preposition, adjective, etc.), or a relationship between hadith

TABLE 12. TF for each tagging in a class using part of speech tagging.

Hadith	POS_Tag	Class	TF		Start with Tag
من روى عني حديثا وهو يرى أنه كذب فهو أحد الكاذبين	من/IN روى/VBD عني/NN حديثا/NN وهو/NNP يرى/VBP أنه/NN كذب/NN فهو/NNP أحد/NNP الكاذبين/DTJJ	Saying	423/595	0.711	IN
ضخم القدمين حسن الوجه لم أر بعده مثله	ضخم/JJ القدمين/DTNNS حسن/NN الوجه/DTNN لم/RP أر/VBP بعده/NNP مثله/NNP	Describing	57/58	0.983	JJ
ثلاث عشرة ركعة منها ثمان ويوتر بثلاث وركعتين بعد الفجر	ثلاث/CD عشرة/CD ركعة/NN منها/NN ثمان/NN ويوتر/NNP بثلاث/NN وركعتين/JJ بعد/NN الفجر/DTNN	Reporting	8/12	0.667	CD
توضأ فغسل وجهه ثلاثا ويديه مرتين مرتين ومسح برأسه فأقبل به وأدبر وغسل رجليه	توضأ/VBP فغسل/NN وجهه/NN ثلاثا/CD ويديه/NN مرتين/NNS مرتين/NNS ومسح/NN برأسه/NN فأقبل/NN به/NN وأدبر/NN وغسل/NN رجليه/NNP	Doing	560/880	0.636	VB

classifications. POS is a technique of assigning a part-of-speech to each word in a sentence; it is useful in information retrieval, text classification, and pre-processing step of parsing. Table 12 shows the Term frequency (TF) for each tagging in the class of hadith using part of speech tagging. From this table we notice that 98% of hadiths that started with adjective (JJ) are related to class describing, 71% of hadiths that started with preposition (IN) are related to class saying, 66% of hadiths that started with cardinal number (CD) are related to class reporting, and finally 63 % of hadiths that started with verb (VB) whether in the past or present tense are related to class doing.

Table 13 shows the relationship between the category of hadith and its main parts such as the verb in base form (VB) or its derivatives (VBD, VBG, VBN, VBP, VBZ), noun (NN), adjective (JJ), and an adverb (RB), using POS-tagging technique, to obtain the parts frequency of hadith, for example number of hadiths that are begins with specific part in the text of hadith, by compute the number of hadiths that begins with this part divided by the sum number of hadiths for each category(Reporting (R), Doing (D), Saying (S), and Describing (De)).

Also, in this study, A random sample of 1600 hadith was selected which divided into four equal subgroups of class A as 25% of each group. The results are summarized in Table 14 shown that 1070 hadith of the sample are classifying as 'Saying' which represent 67%. The next is ''Reporting'' 236 hadith by 14.7%, followed by 'Describing' and 'Doing' as 168(10.5%) and 125(7.8%), respectively. These results conclude that the highest percentage of the 'Sahih' hadith 318(30%) is classifying as ''Saying'', follow by ''Reporting'' which represent 39(16%). Finally, the most untrusted hadith as 'Da'if' and 'Maudu' are classifying as 'Reporting' and 'Describing', respectively.

The chi-Square method is used to test whether an association exists between two categorical variables by comparing

TABLE 13. Relations between the class of hadith and its part of speech tagging.

Start with Tag	No. of Hadith					Percentage			
	R	D	S	De	Total	R	D	S	De
VBP	307	560	203	140	1210	0.254	0.463	0.168	0.116
IN	90	37	423	45	595	0.151	0.062	0.711	0.076
NN	413	371	229	491	1504	0.275	0.2467	0.152	0.326
CD	8	2	2	0	12	0.667	0.166	0.167	0
JJ	1	0	0	57	58	0.017	0	0	0.983
JJR	0	0	3	0	3	0	0	1	0
VBD	477	665	516	817	2475	0.193	0.269	0.201	0.330
VBN	17	5	3	2	27	0.629	0.185	0.111	0.074
NOUN	2	2	7	0	11	0.182	0.182	0.636	0
WP	45	10	49	28	132	0.341	0.076	0.371	0.212
DTNN	84	27	83	40	234	0.359	0.115	0.355	0.171

TABLE 14. Relationships between the classifications of hadith.

		Class A				
		Saying	Reporting	Doing	Describing	Total
Class B	Sahih	318 (30%)	39 (16%)	21 (16.8%)	20 (12%)	398 (25%)
	Hasan	258 (24%)	45 (19%)	75 (60%)	22 (13.1%)	400 (25%)
	Da'if	260 (24%)	73 (31%)	15 (12%)	53 (31.5%)	401 (25%)
	Maudu	234 (22%)	79 (33.5%)	14 (11.2%)	73 (43.5%)	400 (25%)
Total		255 (63.7%)	1070 (67%)	236 (14.7%)	125 (7.8%)	1599 (100%)

the observed values of responses to the values that would be expected, based on the following equation:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{`E_i}$$

where: O_i = the observed frequency. E_i = the expected frequency if O_i a relationship existed between the variables.

Hypotheses:

H_0: Class B (Sahih, Hassan, Da'if, and Maudu) is not related to (associated with) Class A (Saying, Reporting, Doing, and Describing).

H_A: Class B (Sahih, Hassan, Da'if, and Maudu) is related to (associated with) Class A (Saying, Reporting, Doing, and Describing).

From results shown in Table 15, we found that $\chi^2 = 163.879$ and $p < 0.001$; which is a very small probability of the observed data under the null hypothesis of no relationship. So, the null hypothesis is rejected. This concludes that the type of hadith definite by Class B seems to be significantly related to class A at the level of significant $\alpha = 5\%$. This conclusion agrees with the results found in Table 13, (as ex: the most of Sahih hadith is classified as Saying, class).

TABLE 15. Chi-square test.

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	163.879[a]	9	.000
Likelihood Ratio	151.808	9	.001
Linear-by-Linear Association	52.745	1	.000
N of Valid Cases	1599		
*α=0.05			

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 31.11.

TABLE 16. CLASSIFICATION accuracy for each classifier divided by ten folds using cross-validation method.

Classifier	Folds									
	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
LinearSVC	0.92	0.94	0.91	0.94	0.92	0.93	0.93	0.94	0.94	0.91
Logistic regression	0.9	0.94	0.89	0.92	0.92	0.91	0.9	0.92	0.91	0.9
SGD Classifier	0.92	0.94	0.91	0.96	0.92	0.93	0.94	0.93	0.92	0.91

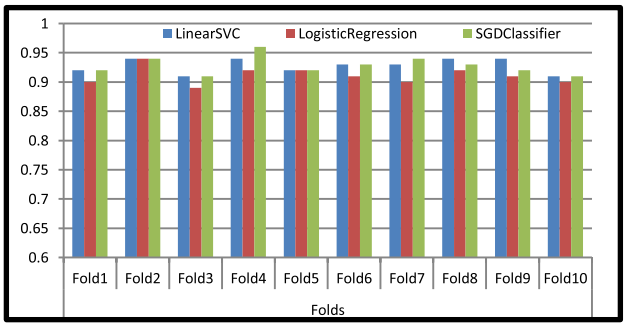


FIGURE 7. The classification accuracy for each classifier divided by ten folds using cross-validation method.

VIII. EXPERIMENTAL RESULTS AND EVALUATION

The Experimental Results and our applications have developed using Python Programming; the results of Hadith classification are evaluated and tested by Python. Different classifier algorithms are used in this study but we reported the best three classifiers of them, which have the highest accuracy whether in the classification of hadith according to what was attributed to the Prophet, according to the Ways of Takhreej the Hadith and the study of Asaneed, or according to the reliability and memory of the reporter based on text of hadith and it's Sanad.

A. EXPERIMENTAL RESULT FOR CLASSIFICATION OF HADITH ACCORDING TO WHAT WAS ATTRIBUTED TO THE PROPHET

In this classification the training data of hadith is divided randomly into ten folds, each fold is held out once, and the classifier is trained on the remaining nine folds (F), to train

TABLE 17. Precision, Recall, and F1-score) for each category using SGDClassifier, Logistic regression, and LinearSVC.

Classifier	SGDClassifier			Logistic regression			LinearSVC		
Class	P	R	F1	P	R	F1	P	R	F1
Describing	0.96	0.96	0.96	0.93	0.95	0.94	0.96	0.96	0.96
Doing	0.95	0.96	0.95	0.95	0.97	0.96	0.95	0.97	0.96
Reporting	0.91	0.92	0.92	0.91	0.87	0.89	0.91	0.91	0.91
Saying	0.9	0.87	0.89	0.86	0.86	0.86	0.9	0.88	0.89
Avg	**0.93**	**0.93**	**0.93**	**0.91**	**0.91**	**0.91**	**0.93**	**0.93**	**0.93**

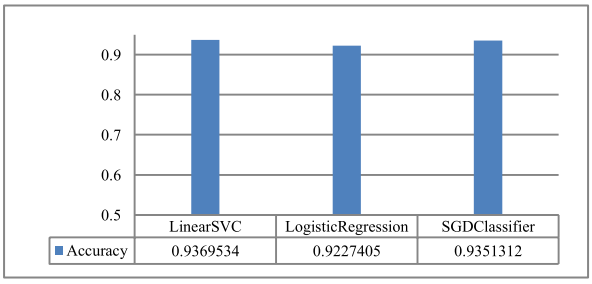


FIGURE 8. Overall percentage accuracies for individual classifier.

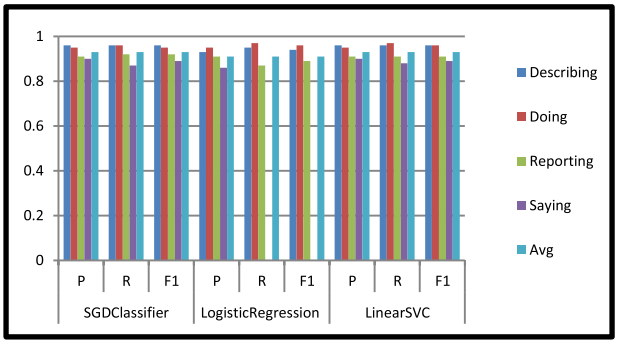


FIGURE 9. Precision, Recall, and F1-score for each category using SGDClassifier, Logistic regression, and LinearSVC.

the classifier on all dataset, the accuracy of each classifier per fold as shown in Table 16, and Figure 7.

In this study, different algorithms have applied, three of them (LinearSVC, Logistic regression, and SGDClassifier) are selected as the best algorithms gave high accuracy, as shown in Figure 8. Each algorithm is tested using cross-validation method, that the best method to evaluate the algorithm as the previous that mentioned before in section 1 [28], [32] We note from the experiments results that the accuracy of the classification of hadith according to what was attributed to the Prophet is better than the results of classification of the hadith according to the reliability and memory of the reporters, because the first classification only depends on the text of hadith that contains some features and keywords that distinguishing each class, therefore; there is a relationship between the content of hadith text and its class.

Table 17 and Figure 9 show the F1-measure (F1), and Recall (R) and Precision (P) for the individual category. These categories are Describing, Doing, Reporting, and Saying,

		Corrected Class		
Predicted Class	Class	A	B	C
	A	TP_a	F_ba	F_ca
	B	F_ab	TP_b	F_cb
	C	F_ac	F_bc	TP_c

FIGURE 10. Confusion matrix for different classes.

according to what was attributed to the Prophet using cross-validation method. We note from the values of F1, R and P there no bias between these values. These values are close to each other for each class, which indicate the efficiency of the classifier for each individual category and the model is more generalization and avoiding to overfitting and underfitting problem.

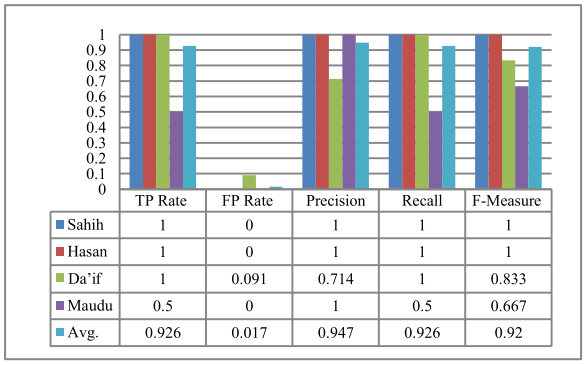
B. EXPERIMENTAL RESULTS ACCORDING TO THE METHODS OF TAKHREEJ THE HADITH AND THE STUDY OF ASANEED

A Decision tree algorithm is applied to predict the class of hadith according to the Methods of Takhreej the Hadith and the study of Asaneed. It is one of the predictive modeling approaches used in statistics, data mining and machine learning, to go from an observation about items represented in branches to conclusions represented in the leaves, this algorithm was tested by cross-validation method.

The performance of this algorithm has been measured using many measures: mainly: Precision, Recall, and F-measure according to equation 12, equation 13 and equation 14 respectively, as shown in figure11, based on the confusion matrix. To understand the performance of this algorithm it is necessary to understand the confusion matrix as shown in Figure 10.

Figure 10 shows the confusion matrix for a different class classification. TP_a, TP_b, and TP_c are correct classifications, but the rest of matrix is incorrectly predicted class, where the TP_a, TP_b, and TP_c represented the number of samples that are correctly classified from class A, B, and C respectively. F_ab, F_ac represented the samples from class A that were incorrectly classified as class B and class C, the F_ba, F_bc represented the samples from class B that were incorrectly classified as class A and class C, the F_ca, F_cb represented the samples from class C that were incorrectly classified as class A and class B, as will show at experimental results according to the Ways of Takhreej the Hadith and the study of Asaneed.

A confusion matrix is a matrix form that is used to evaluate the performance of a classifier (algorithm) on a set of training and testing data, and to determine the number of hadiths that are correctly specified to the given class or the number of hadith which are incorrectly defined to the class. From this matrix, we can know the no of hadith that belong to TP, TN, FP or FN for each class. Each row of the matrix represents the number of instances in a predicted class while each column represents the number of instances in an actual class.



	TP Rate	FP Rate	Precision	Recall	F-Measure
Sahih	1	0	1	1	1
Hasan	1	0	1	1	1
Da'if	1	0.091	0.714	1	0.833
Maudu	0.5	0	1	0.5	0.667
Avg.	0.926	0.017	0.947	0.926	0.92

FIGURE 11. Classification report for each category using Decision tree classifier.

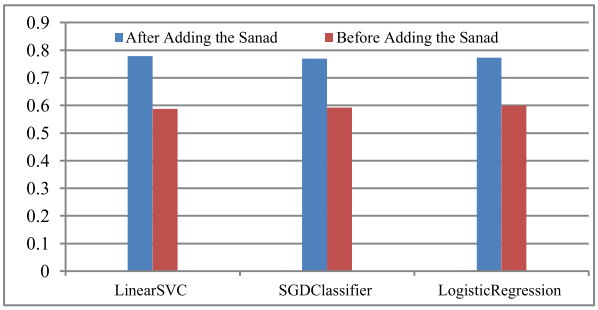


FIGURE 12. Overall percentage accuracies for individual classifier.

TABLE 18. OVERALL percentage accuracies for individual classifier after and before adding the Sanad.

Classifier	After Adding the Sanad	Before Adding the Sanad	Increasing Accuracy
LinearSVC	0.778864971	0.588177988	0.190686982
SGDClassifier	0.769569424	0.592788863	0.176780561
LogisticRegression	0.772994129	0.599412985	0.173581144

C. EXPERIMENTAL RESULTS ACCORDING TO THE RELIABILITY AND MEMORY OF THE REPORTER BASED ON THE TEXT OF HADITH AND ITS SANAD

In this classification the training data of hadith is divided randomly into ten folds, each fold of them is held out once to test the classifier and the classifier is trained on the remaining nine folds and so on in each one of the ten, to train the classifier on all dataset. Figure 12 shows the overall accuracy for the best three classifiers, each classifier of them was tested using cross-validation method as the previous that mentioned before in section 7.1. We note from the results of the experiment that the accuracy of the hadith classification according to the Reliability and Memory of the Reporters has been increased when we added the Sanad to the text of hadith as shown in Table 18 because this classification depends mainly on the Sanad. Therefore, there is no explicit relationship between content of hadith text and its class.

Table 19 and Figure 13 show the F1-measure (F1), and Recall (R) and Precision (P) for the individual category. These categories are Sahih, Hasan, Da'if, and Maudu, according

TABLE 19. F1-score, Recall, and Precision for each category using SGDClassifier, Logistic regression, and LinearSVC.

Classifier	SGDClassifier			Logistic regression			LinearSVC		
Class	P	R	F1	P	R	F1	P	R	F1
Da'if	0.82	0.87	0.84	0.82	0.86	0.84	0.83	0.87	0.85
Hassan	0.69	0.64	0.67	0.72	0.62	0.67	0.71	0.65	0.68
Maudu	0.85	0.84	0.84	0.83	0.85	0.84	0.84	0.85	0.85
Sahih	0.74	0.75	0.74	0.73	0.78	0.75	0.74	0.77	0.76
Avg.	0.77	0.77	0.77	0.77	0.77	0.77	0.78	0.78	0.78

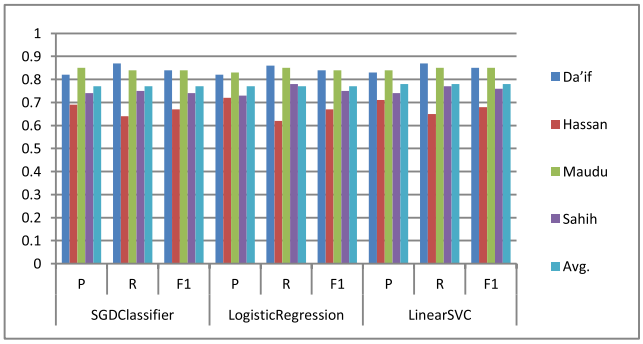


FIGURE 13. F1-score, Recall, and Precision for each category using SGDClassifier, Logistic regression, and LinearSVC.

to the reliability and memory of the reporters using cross-validation method.

IX. CONCLUSIONS

The Hadith is the second source of Islam after the Qur'an and the fundamental resource of legislation in the Islamic community. Therefore; in this study, three methodologies are applied to classify the hadith automatically into different categories according to the Reliability and Memory of the Reporters and based on the Sanad of hadith using machine learning techniques. The Experimental results revealed that adding the Sanad of hadith to the text in the classification process have a significant impact on the increase the classification accuracy because this classification depends on the Sanad of hadith, which identifying and evaluating the hadith degree. We note from the experiments results that the accuracy of the classification of hadith according to what was attributed to the Prophet is better than the results of classification of the hadith according to the Reliability and Memory of the Reporters, because the first is depends on the content of the text, based on it contains some features and keywords that distinguishing each class, so there is relation between the content of the hadith text and its class, this relation makes the classifier is able to recognize and distinguish between the classes of hadith. Based on Experimental results, obviously there is a close relationship between the classification of hadith according to the reliability and memory of the reporter and according to what was attributed to the Prophet. Finally, the experimental results showed that the best three classi-

fiers gave high accuracy whether in classifying the hadith according to the Reliability and Memory of the Reporters or according to what was attributed to the Prophet are LinearSVC; it achieved accuracy reached up to 93.69%, 77.88, followed by SGDClassifier, reached up to 93.51%, 76.95, and finally, Logistic regression reached up to 92.27%, 77.29. These classifiers gave high accuracy compared to other classifiers, because they are able to identify the essential features relevant to target data in the training stage, and they can deal with missing values.

REFERENCES

[1] T. Ismail, R. Baru, A. Hassan, and A. Salleh, "The matan and sanad criticisms in evaluating the hadith," *Asian Social Sci.*, vol. 10, no. 21, 2014.
[2] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Urbana, IL, USA: Univ. Illinois Urbana Champaign, 2012. [Online]. Available: http://www.mkp.com
[3] N. Mohammed Al-Kabi, A. Heider Wahsheh, M. Izzat Alsmadi, and A. M. A. Al-Akhras, "Extended topical classification of hadith arabic text," *Int. J. Islamic Appl. Comput. Sci. Technol.*, vol. 3, no. 3, pp. 13–23, Sep. 2015.
[4] K. Jbara, "Knowledge discovery in Al-Hadith using text classification algorithm," *J. Amer. Sci.*, vol. 6, no. 11, pp. 409–419, Jan. 2010.
[5] K. Bilal and S. Mohsen, "Muhadith: A cloud based distributed expert system for classification of ahadith," in *Proc. 10th Int. Conf. Frontiers Inf. Technol.*, Islamabad, India, Dec. 2012, pp. 73–78.
[6] M. M. Najeeb, "Towards innovative system for hadith Isnad processing," *Int. J. Comput. Trends Technol.*, vol. 18, no. 6, pp. 257–259, Dec. 2014.
[7] K. A. Aldhlan, A. M. Zeki, A. M. Zeki, and H. A. Alreshidi, "Novel mechanism to improve hadith classifier performance," in *Proc. Int. Conf. Adv. Comput. Sci. Appl. Technol. (ACSAT)*, Nov. 2012, pp. 512–517.
[8] M. A. Saloot, N. Idris, R. Mahmud, S. Jaafar, D. Thorleuchter, and A. Gani, "Hadith data mining and classification: A comparative analysis," *Artif. Intell. Rev.*, vol. 46, no. 1, pp. 113–128, Jun. 2016.
[9] K. Aldhlan, A. Zeki, and H. Alreshidi, "Improving knowledge extraction of Hadith classifier using decision tree algorithm," in *Proc. Int. Conf. Inf. Retr. Knowl. Manage.*, Mar. 2012, pp. 148–152.
[10] M. Tahan, *Asoul Al-Takhreej and Study of Asaneed*. Riyadh, Saudi Arabia: Al-Maref, 1996.
[11] A. A. S. Al-Luhaidan, "Scientific methods in the graduation of Hadith," College Fundam. Religion Riyadh, Imam Muhammad Bin Saud Islamic Univ., Riyadh, Saudi Arabia, Tech. Rep., 1999.
[12] (1998). *Graduation and Studying the Al-Asaneed*. [Online]. Available: http://www.ahlalhdeeth.com
[13] *Methods of Graduation of the Hadith of the Messenger of Allah (Peace and Blessings of Allah be Upon Him)*, Al-Azhar Univ., Cairo, Egypt, pp. 1799–1987, 1987.
[14] *Mustalah Al-Hadith*. [Online]. Available: http://ia802908.us.archive.org/26/items/Marthads_Library/Mustalah.simple.pdf and https://www.archive.org/download/Marthads_Library/Mustalah.simple.pdf
[15] Abdul Ghani Al-Tamimi. (2011). *Graduation of the Prophet's Hadith*. [Online]. Available: https://d1.islamhouse.com/data/ar/ih_books/../ar_tkhreeg_Al_7deeth_An_Nbwee.doc
[16] S. Hassan. *Introduction to the Science of Hadith Classification*. Accessed: Feb. 2019. [Online]. Available: http://www.ahya.org/modules.php?name=Sections&op=viewarticle&artid=7 and https://www.ahya.org
[17] H. I. Witten, E. Frank, and A. M. Hall, *Data Mining Practical Machine Learning Tools and Techniques*, 2nd ed. San Francisco, CA, USA: Morgan & Claypool, 2005.
[18] A. Abdullah, G. Tan, A. Khaled, and H. Rajeh, "The effect of preprocessing on arabic document categorization," *Algorithms*, vol. 9, no. 2, p. 27, Apr. 2016.
[19] H. M. Abdelaal, A. N. Elmahdy, A. A. Halawa, and H. A. Youness, "Improve the automatic classification accuracy for Arabic tweets using ensemble methods," *J. Elect. Syst. Inf. Technol.*, vol. 5, no. 3, pp. 363–370, Dec. 2018.
[20] Y. A. Alhaj, J. Xiang, D. Zhao, M. A. A. Al-Qaness, M. A. Elaziz, and A. Dahou, "A study of the effects of stemming strategies on arabic document classification," *IEEE Access*, vol. 7, pp. 32664–32671, 2019.
[21] Z. Qiu, C. Gurrin, A. R. Doherty, and A. F. Smeaton, "Term weighting approaches for mining significant locations from personal location logs," in *Proc. 10th IEEE Int. Conf. Comput. Inf. Technol.*, Bradford, U.K., Jun./Jul. 2010, pp. 20–25.
[22] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manage.*, vol. 24, no. 5, pp. 513–523, Jan. 1988, doi: 10.1016/0306-4573(88)90021-0.
[23] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *J. Mach. Learn. Res.*, vol. 3, pp. 1289–1305, Mar. 2003.
[24] M. Saad and W. Ashour, "Arabic text classification using decision trees," in *Proc. 12th Int. Workshop Comput. Sci. Inf. Technol. (CSIT)*, Moscow, Russia, vol. 2, 2010.
[25] S. Teufel, "Term weighting and the vector space model," Natural Lang. Inf. Process. Group, Inf. Retr. Comput. Sci. Tripos Part II, Univ. Cambridge, Cambridge, U.K., Tech. Rep.
[26] S. Liu, W. Bai, N. Zeng, and S. Wang, "A fast fractal based compression for MRI images," *IEEE Access*, vol. 7, pp. 62412–62420, 2019.
[27] S. Baraa, O. Nazlia, and S. Zeyad, "An automated arabic text categorization based on the frequency ratio accumulation," *Int. Arab J. Inf. Technol.*, vol. 11, no. 2, pp.–221, Mar. 2014.
[28] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Urbana, IL, USA: Univ. Illinois Urbana Champaign, 2012. [Online]. Available: http://www.mkp.com
[29] *Data Mining Tutorial, Data Pattern Evaluation, Simply Easy Learning by Tutorialspoint.com*. [Online]. Available: https://www.tutorialspoint.com/data_mining/data_mining_pdf_version.htm
[30] A. Mahmoud, H. Khan, Z. Rehman, and W. Khan, "Query based information retrieval and knowledge extraction using Hadith datasets," in *Proc. 13th Int. Conf. Emerg. Technol. (ICET)* Dec. 2017, pp. 1–6. [Online]. Available: https://www.researchgate.net/publication/323193943
[31] S. Bassinet, A. Madani, M. Al-Sarem, and M. Kissi, "Feature selection using an improved Chi-square for Arabic text classification," *J. King Saud Univ., Comput. Inf. Sci.*, to be published.
[32] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 2nd ed. Urbana, 8IL, USA: Univ. Illinois Urbana Champaign, 2006.

HAMMAM M. ABDELAAL received the B.Sc. and M.Sc. degrees in computers and systems engineering from Al-Azhar University, Egypt. He is currently pursuing the Ph.D. degree in computers engineering with Minia University. He is a Researcher with Minia University and also an Assistant Lecturer with the Higher Institute for Computers and Information Systems. His main areas of research interests are machine learning techniques, supervised learning algorithms, natural language processing, and data mining.

ABDELMOTY M. AHMED received the B.S. and M.S. degrees in systems and computers engineering from the Faculty of Engineering, Al-Azhar University, Cairo, Egypt, in 2003 and 2008, respectively. He is currently pursuing the Ph.D. degree in systems and computer engineering. He is also a Senior Lecturer with the Computer Engineering Department, College of Computer Science, King Khalid University, Abha, Saudi Arabia. His research interests are image processing, biometrics, artificial intelligence, pattern recognition, machine learning, and computer vision. He is also interested in researching the technical fields that serve deaf and dumb and also works in the automatic translation of the Arabic Sign Language.

WADE GHRIBI received the B.Sc. and M.Sc. degrees in computer science and engineering from the Department of Computer Systems and Networks, National Aerospace University Kharkov, Ukraine, in 2003, and the Ph.D. degree from the Computer Engineering Faculty, Kharkov National University of Radio Electronics, Kharkov, Ukraine, in 2007. He is currently working as an Assistant Professor with the College of Computer Science, King Khalid University, Abha, and KSA. His research interests include digital system design and digital sign, and image processing.

HASSAN A. YOUNESS ALANSARY received the B.Sc. and M.Sc. degrees from Assiut University, Assiut, Egypt, and the Ph.D. degree from the Graduate School of Information Science and Technology, Osaka University, Japan, with the cooperation of Ain Shams University, Egypt. He worked for IBM Company and Mentor Graphics, Egypt. He is currently an Associate Professor with Minia University, and also the Chairman of the Computers and Systems Engineering Department. His research interests include integrated system design, fault tolerance, HW/SW co-design, parallel computers, embedded systems, GPGPU, APU and MPSoCs, and homogeneous/heterogeneous systems.