The Relationship between different Classifications of Hadiths Based on rules Mining

Hammam Abdelaal[1], Ahmed Elkattan[2], Hussein Elshafie[3] and Abdelmoty Ahmed[4]

[1]Department of Information Technology, Faculty of Computers and information, Luxor University, Egypt
[2]Faculty of Economics and Business, Complutense University of Madrid, Spain
[3]Department of Computer Science, Faculty of Computers and information, Luxor University, Egypt
[4]College of Computer Science, Nahda University, Beni Suef, Egypt

Abstract- There are many classifications of Hadith mainly: according to reliability and memory of hadith' narrators (sahih, hasan, maudu, da'if), and according to the sayings, actions or characteristics of the Prophet (saying, doing, reporting, describing), each Hadith contains three main parts: Sanad, Matn and Taraf. In this study we try to find the relationship between the different classifications of hadith using Association rule mining, part of speech (POS) and Chi-square techniques, based on the text and Sanad of hadith to build a model that is able to distinguish and categorize the hadith categories, using supervised learning. The experimental results showed that the relation (confidence) between saying & sahih is 79.99%, doing & hasan is 64.5%, reporting & da'if is 64.84, describing & maudu is 58.5%, and the best classifier has given high accuracy been Naïve bayes NB; it achieved higher accuracy reached up to 97.5 %, followed by the LinearSVC and K- Neighbors reached up to 96.5%.

Keywords—Data Mining, Learning Algorithms, Association Rules mining, POS, Chi-square, Hadith Text

1. INTRODUCTION

Mining of data is extracting knowledge from large amounts of data; it is a process of analyzing a large number of data to find a logical relationship summarizes data in a new way is understandable and useful to the owner of the data [1] data mining is divided into two types: descriptive and predictive data mining. Two of the most important techniques related to predictive and descriptive data mining are the classification process and association rules [2], these techniques are used in this study to classify the hadith into different categories and find a relation between these classifications. In this study, the classification technique is used, based on supervised learning algorithms to classify the hadiths into different categories according to what was attributed to the Prophet Muhammad mainly: Saying, Doing, Describing, and Reporting, additional to finding a relationship between these classifications. Prophetic hadith is what was added to the Prophet Muhammad, it includes the sayings, actions, reporting, or characteristics of the Prophet, each Hadith consists of three parts: the text known as the Matn, the Sanad (narrators' chain of hadith) [3] [4] and the Taraf (beginning sentence of the text that refers to the sayings, characteristics or actions of the Prophet, or his concurrence with others action). In this study, the dataset of hadith is split using the cross-validation method. In this method the dataset of hadith is divided randomly into a number of 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 [5].

Hadith classification depends mainly on the Sanad; it is a 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 criteria which through the validity of the Hadith is judged, and based on the definition of each class like Sahih, Da'if [7] [8]. 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 narrator, and etc. These conditions and criteria were formulated to a set of features to train the learning algorithm for recognize the class of hadith; each feature of them has a set of possible values which represent the dataset of hadith as shown in Table1. Also every hadith - except 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 [9] [10]

2. PROPOSED SYSTEM

The model is consists of different stages mainly: the first one is hadith collection from different sources according to the classifications of each hadith, and the text pre-processing, followed by text representation, then feature selection and finding the relation between text words[19][22], after that the learning classifier to build the model, and finally evaluating and testing the classifier performance using unseen examples of hadith, additional to finding a relationship between these classifications according to the rate of confidence between these classification using association rule mining

## 3. CLASSIFICATION OF HADITH

In this Classification, the hadith has been classified into different categories mainly: Describing, Saying, Reporting and Doing according to what was attributed to the Prophet. Hadith consists of two main parts, the most important of which is Sanad and the Matn (Text). In this classification, the hadith has been classified based on the Hadith text only, and the Sanad is ignored, it irrelevant of hadith categories. Hadith text has been processed as like other Arabic texts rest to convert it to a form that is suitable for classification process [10] [11], through these steps mainly: Remove the Sanad, Normalization, Tokenization, Stop Word Removal, and the Stemming [12][19]. After that Term Weighting to convert the full text into vectors such as index terms, which frequently used in information retrieval, indexing and text classification [13] [14] and determine the most important words in a documents collection for classification purpose by calculate the term weighting for each word in a document, then the Feature Selection to removing undesirable data, which helps us to select the essential features that are relevant to target data, and ignore the data that is irrelevant to the target data, in additional to enable the learning algorithm to train faster and improve the classification accuracy [16], finally learning and evaluating the classifier

4. RELATION BETWEEN CLASSIFICATIONS

In this study, we try to discover a relationship between the classification of hadith according to what was issued from Prophet Muhammad and according to the Reliability and Memory of the Reporters through hadith classifications using association rules technique. After building different classifier models using supervised learning algorithms according to what was attributed to the Prophet, the best classifier given high accuracy has applied to test 1600 hadith randomly, which divided into 400 samples per category from the dataset of hadith that belong to the classification of hadith according to the reliability and memory of the reporters, and the output of the classifier as shown in table 1 and figure 2, to find the support count and confidence between these classes using equation 8, and equation 9 [15]. According to table 1, we can find a relation between these classes using association rule mining by computing the support count and confidence between these classes to discover dependency rules that will predict the presence of a class in a dataset of hadith based on the presence of other classes.

$$Support\ (X{-}{-}{-}{>}Y) = P\ (XUY)$$
$$Confidence\ (X{-}{-}{-}{>}Y) = P\ (X|Y)$$
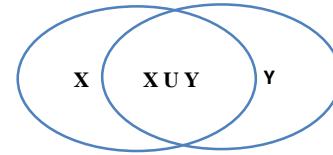
The Rules of (X--->Y) are

$$Support = \frac{frq(X,Y)}{N}$$

$$Confidence = \frac{frq(X,Y)}{frq(X)}$$

Confidence or certainty equal to 50% means that if a Hadith is Sahih, there is a 50% chance that the hadith is the category of saying. A 50% support means that 10% of all of the datasets showed that of a Sahih category and a saying category were there together.

Table 1 Relationship between hadith categories

Class	Da'if	Hasan	Maudu	Sahih	Total
Saying	260	258	234	318	1070
Doing	15	75	14	21	125
Reporting	73	45	79	39	236
Describing	53	22	73	20	168
Total	401	400	400	398	1599
Confidence	64.84	64.5	58.5	79.99	66.92

From table 1, 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 Describing, and finally the Doing class. Then the large number of hadiths that belong to saying class found in Sahih, followed by Da'if, Hasan, and finally the Maudu class as shown in figure 1.
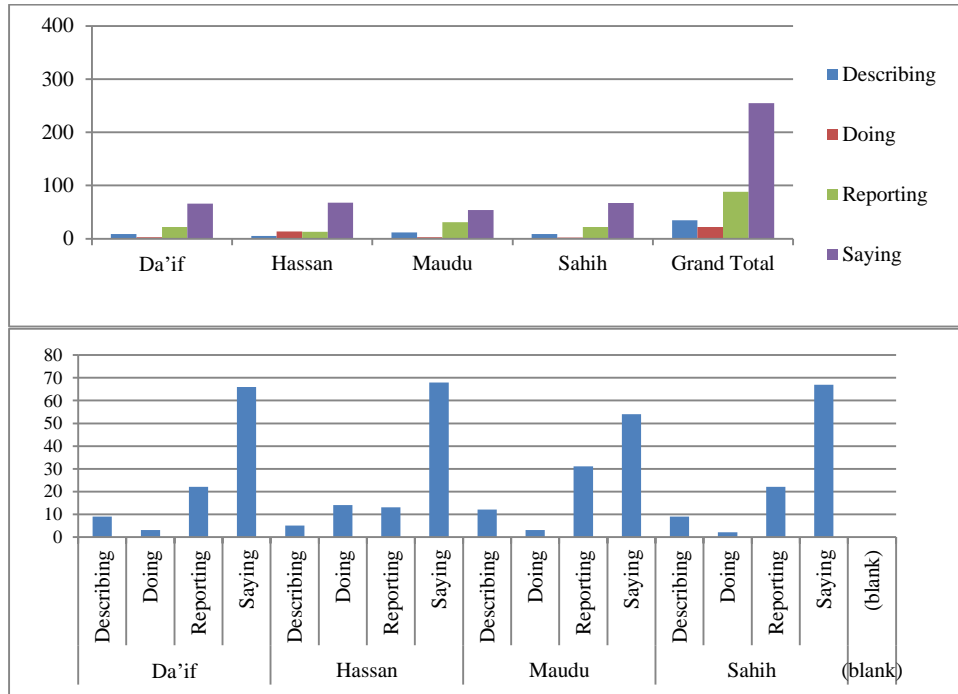
Figure 1 Number of Hadith in each Class

- Confidence (Sahih, saying) = P (318/398) = 79.99%, Confidence (Da'if, saying) = P (260/401) = 64.84%

- Confidence (Maudu, doing) = P (14/400) = 3.5%, Confidence (Sahih, describing) = P (20/398) = 5.02%

- Confidence (saying, sahih) = P (318/1070) = 29.72%, Confidence (saying, da'if) = P (260/1070) = 24.29%

- Confidence (doing, maudu) = P (14/125) = 11.2%, Confidence (describing, sahih) = P (20/168) = 11.90%

- (Reporting, Describing & hasan)➜ 47.18 %, (Reporting, Describing & maudu) ➜ 91.57%

- Support (Sahih, describing) = P (20/1599) = 1.52%, Support (Sahih, saying) = P (318/1599) = 19.89%

From these results, we conclude that 79.99% of *Sahih* hadiths are belong to *saying* class, and 11.2 % of *maudu* hadiths are belong to *doing* class, also 1.52% of *Sahih* hadiths are *describing* class and so on. One of them is 29.72 % of *saying* hadiths are belong to *Sahih* class, and 11.9% of *describing* hadiths are belong to Sahih class, also 5.02% of Sahih hadiths are describing class. Also, most the *da'if* and *maudu* hadiths are belong to *reporting* and *describing* categories.

5. POS TAGGING TECHNIQUE

POS tagging is a technique of determine a part-of-speech to each word in a sentence, that have similar grammatical properties, such as noun, verb, adjective, adverb, conjunction, pronoun, preposition, numeral, interjection, each part of them determine with key word as shown in the

below figure, it useful in information retrieval, text classification, and preprocessing step of parsing [21]. Table 2 shows the number of occurrences of each Part (JJ, CD, VB... etc.) of text in each category and the number of times each part appears at the beginning of hadith text for each category to find a relationship between these categories

Tag	Description
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker

Tag	Description
PRP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	to
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense

Table 2 part of speech for sample of hadiths

Hadith	POS_Tag	Class	TF		Start 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	0.710924	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	0.982759	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	0.666667	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	0.636364	VB

Start with	Reporting	Doing	Saying	Describing	Total
VBP	307	560	203	140	1210
IN	90	37	423	45	595
NN	413	371	229	491	1504
VBD	477	665	516	817	2475
NOUN	2	2	7	0	11
WP	45	10	49	28	132
DTNN	84	27	83	40	234

6. CHI-SQUARE TESTING TECHNIQUE

The study adopted the Chi-square testing technique to test the association between classification dismissing-including narrators, text of hadith, and reference—and class. The results revealed that all classification dimensions positively affect class. Table 3 provides a summary of results for correlation coefficients between classification dimensions and class. The results confirmed a positive and significant relationship between the narrator and class (χ^2 = 36.000, df = 9, p = <.001), indicating a strong association (Phi =.750, Cramer's V = .433, p = <.001). A similar pattern of results was observed for the relationship between reference and class (χ^2 = 91.429, df = 9, p = <.001), highlighting a strong relationship between the two binary variables (Phi =.795, Cramer's V = .690, p = <.001). In contrast, results showed an insignificant relationship between the text of hadith and class (χ^2 = 2.333, df = 9, p = .985).

Table 3 Correlation between classification dimensions and class

Dimension correlation	Chi-square	Value	df	Asymp. Sig. (2-sided)	Effect size	Value	Approximate Significance
Narrators and class	Pearson Chi-Square	36.000[a]	9	<.001	Phi	.750	<.001
	Likelihood ratio	37.877	9	<.001	Cramer's V	.433	<.001
	Linear-by-linear association	9.998	1	.002			
Text of hadith and class	Pearson Chi-Square	2.333[a]	9	.985	Phi	.191	.985
	Likelihood ratio	2.262	9	.987	Cramer's V	.110	.985
	Linear-by-linear association	.319	1	.572			
Reference and class	Pearson Chi-Square	91.429[b]	9	<.001	Phi	.795	<.001
	Likelihood ratio	98.991	9	<.001	Cramer's V	.690	<.001
	Linear-by-linear association	40.443	1	<.001			

N = 64.

a = 12 cells (75.0%) have expected count less than 5. The minimum expected count is 2.00.

b = 13 cells (81.3%) have expected count less than 5. The minimum expected count is 1.25.

7. EXPERIMENTAL RESULTS

Different algorithms have used in this study, but we reported the best three algorithms, which have the highest accuracy. In this classification the dataset of hadith is divided randomly into ten parts, each part is held out once to test the classifier, and the classifier is trained on the remaining nine parts to train and test all dataset known as cross validation method [17][18]. Table 2 and Figure 2 show the Precision, Recall, and F1-measure (F1) for individual category. These categories are Describing, Doing, Reporting, and Saying, according to what was attributed to the Prophet using cross-validation method. We note from the values of Precision, Recall, and F1 there no bias between these values, which indicate the efficiency of the classifier for each individual category and the model is more generalization.

Table 2 Precision, Recall, and F1-score) for each category using K-Neighbors, SVC, and NB

Classifier	K-Neighbors			SVC			NB		
Class	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Describing	0.94	0.98	0.96	0.92	0.99	0.95	0.98	0.99	0.98
Doing	0.97	0.99	0.98	0.99	1	0.99	0.97	0.99	0.98
Reporting	0.98	0.91	0.94	0.99	0.89	0.94	0.99	0.92	0.95
Saying	1	0.99	0.99	0.99	0.99	0.99	0.89	0.99	0.94
Avg	0.97	0.96	0.96	0.97	0.97	0.97	0.97	0.97	0.97
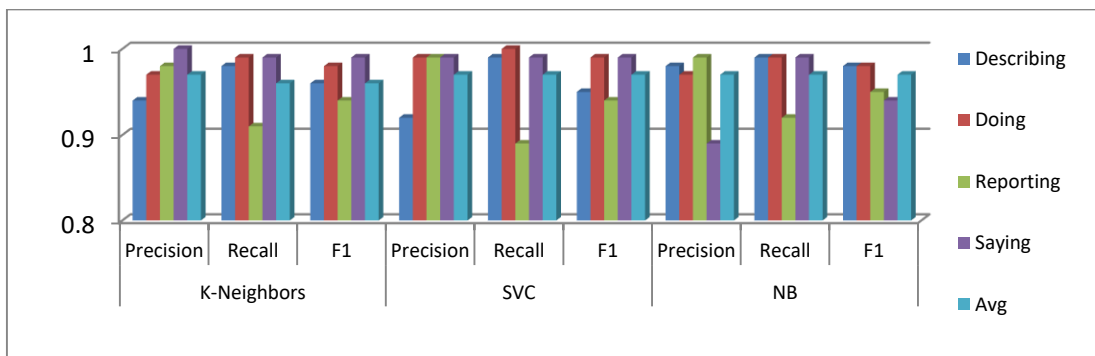
Figure 2 Precision, Recall, and F1-score for each category using K-Neighbors, SVC, and NB

6. CONCLUSIONS

There are relationship between the classifications of hadith. This relation makes the classifier is able to recognize and distinguish between the hadith classes. Most of these hadiths that related to Reliability and Memory of the Reporters are classified as saying class, followed by Reporting, then the Describing, and Doing as the following (saying & sahih is 79.99%, doing & hasan is 64.5%, reporting & da'if is 64.84, describing & maudu is 58.5%). In addition to that, the highest percentage of the Sahih hadiths was classifying as Saying, followed by Reporting. Finally, the experimental results showed that the best classifier had given high accuracy was NB; it achieved higher accuracy reached up to 97.5 % followed by SVC, reached up to 96.5, and finally, K-Neighbors reached up to 96.45.

References

1. Han, J., 'Data Mining Concepts and Techniques', Second Edition, Morgan Kaufmann Publishers, San Francisco, (2006).
2. Mahmoud, H. Khan2, Z. Rehman1, and W Khan, Query-based information retrieval and knowledge extraction using Hadith datasets, Conference Paper · December 2017, https://www.researchgate.net/publication/323193943.
3. Tuan Ismail, R. Baru2, A. Hassan, and A. Salleh3, 'The Matan and Sanad Criticisms in Evaluating the Hadith,' Asian Social Science; Published by Canadian Center of Science and Education, Vol. 10, No. 21; 2014.
4. HM. ABDELAAL and H A. YOUNESS, 'Hadith Classification using Machine Learning, Techniques According to its Reliability', ROMANIAN JOURNAL OF INFORMATION SCIENCE AND TECHNOLOGY, Volume 22, Number 3-4, 2019.
5. Jiawei Han, Michelins Kamber, and Jian Pei, "Data Mining: Concepts and Techniques," the Third Edition University of Illinois at Urbana Champaign, 2012 Elsevier Inc. Web site at www.mkp.com.
6. Graduation and studying the Al-asaneed, To Sheikh Dr. Hattem B. Al Sharif, published by forum Ahl al-hadith, www.ahlalhdeeth.com, Summer-1419, 1998.
7. Methods of graduation of the hadith of the Messenger of Allah (peace and blessings of Allah be upon him), Abo Mohamed Abd Elhady- Department of Hadith - Al-Azhar University, 1799-1987.
8. Mustalah al-hadith http://ia802908.us.archive.org/26/items/Marthads_Library/Mustalah.simple.pdf, book name Mustalah Simple'
9. Abdul Ghani Al-Tamimi, Graduation of the Prophet's Hadith, www.ktibac.com , 2011 https://d1.islamhouse.com/data/ar/ih_books/.../ar_tkhreeg_Al_7deeth_An_Nbwee.doc
10. Abdullah A, Guanzheng TAN, Khaled A, and H Rajeh 'The Effect of Pre-processing on Arabic Document Categorization', Algorithms, Published: 18 April 2016.
11. HM. ABDELAAL and H A. YOUNESS, 'Improve the automatic classification accuracy for Arabic tweets using ensemble methods', Journal of Electrical Systems and Information Technology, published by Elsevier, December 2018.
12. Y. Alhaj, J. XIANG, 'A Study of the Effects of Stemming Strategies on Arabic Document Classification', IEEE Access journal, February, 2019.

13. Salton, G.; Buckley, C. Term-weighting approaches in automatic text retrieval. Inf. Proc. Manga. 1988, 24, 513–523, http://dx.doi.org/10.1016/0306-4573 (88)90021-0.

14. Y. Alhaj, Wiraj Udara, and H M. ABDELAAL, "Efficient Feature Representation Based on the Effect of Words Frequency for Arabic Documents Classification", Conference Paper: November 2018.

15. Jiawei Han, Michelins Kamber, and Jian Pei, "Data Mining: Concepts and Techniques," the Third Edition University of Illinois at Urbana Champaign, 2012 Elsevier Inc. Web site at www.mkp.com.

16. HM. abdelaal, Abdelmoty M. Ahmed and Hassan A. Youness, "Knowledge Discovery in the Hadith according to the reliability and memory of the reporters using Machine learning techniques" IEEE Access, DOI: 10.1109/ACCESS.2019.2944118, 9 -2019.

17. Abdullah A, Guanzheng TAN, Khaled A, and H Rajeh 'The Effect of Pre-processing on Arabic Document Categorization', Algorithms, Published: 18 April 2016.

18. HA AbdElghfar, AM Ahmed, AA Alani, and HM AbdElaal, "A model for Qur'anic sign language recognition based on deep learning algorithms", Journal of Sensors, 2023

19. Yousif A. Alhaj, Wiraj Udara, "Efficient Feature Representation Based on the Effect of Words Frequency for Arabic Documents Classification", Conference Paper: November 2018.

20. S. Wang, D. Li, X. Song, Y. Wei, and H. Li, ''A feature selection method based on improved Fisher's discriminant ratio for text sentiment classification,'' Expert Syst. Appl., vol. 38, no. 7, pp. 8696–8702, 2011

21. J Awwalu, S Abdullahi, and A Evwiekpaefe, "PARTS OF SPEECH TAGGING: A REVIEW OF TECHNIQUE", DOI:10.33003/fjs-2020-0402-325, Journal of Sciences 4(2):712-721, June 2020

22. HA AbdElghfar AM Ahmed, AA Alani, and HM AbdElaal, "QSLRS-CNN: Qur'anic sign language recognition system based on convolutional neural networks", the Imaging Science Journal 72 (2), 254-266, 2022