The Effect of Clustering Classification and Pre-processing of Text on Improving the Accuracy of Hadith

Prepared by

Dr. Berihan R. Elemary

Statistics & Insurance Dep. Faculty of Commerce, Damietta University.
Faculty of Business and Economics,
Badr University (BUC in Cairo)

berihanelemary@gmail.com

Scientific Journal for Financial and Commercial Studies and Researches (SJFCSR)

Faculty of Commerce – Damietta University

Vol.2, No.1, Part 1., Jan. 2021

APA Citation:

Elemary, B. R. (2021). The Effect of Clustering Classification and Pre-processing of Text on Improving the Accuracy of Hadith**.** ***Scientific Journal for Financial and Commercial Studies and Researches (SJFCSR),*** Vol.2 (1) Part1. pp.549- 576.
Website: https://cfdj.journals.ekb.eg/

Abstract

Clustering is widely used in the context of the text, especially in classification. On the other hand, text pre-processing has a major impact on improving accuracy. This study aims to study of effect text pre-processing on improving the accuracy of hadith text, and building a model to classify the hadith categories into Saying, Doing, Reporting, and Describing, according to what was attributed to the Prophet Muhammad (PBUH), using learning algorithms. Also, this study is concerned with studying the impact on improving the accuracy of classification on text on different classifications of Hadith according to the text of Hadith and four categories of accuracy. Two Way Cluster Analysis was used to classify Hadiths according to the tags present in each Hadith. Results showed that the number of tags could be used to distinguish different classes of Hadith especially the Da'if and Maudu as the classification results showed distinct groups of both types while classification showed low accuracy or capability of the separation between Sahih and Hasan Hadiths. Also, the experimental results showed that the best classifier had given high accuracy was SGD Classifier; it achieved higher accuracy followed by Logistic Regression, and finally, Bernoulli NB.

Key Words: Clustering; Text Classification; Learning algorithms; Text Pre-processing; Hadith.

1. Introduction

Over long periods of time, Islamic documents were investigated and published to be a documented reference and could be applied in the different life activities according to the teachings of the Islamic religion. Many kinds of research were carried to introduce and extend techniques in processing these documents that will make them more simplify for E-users to group and identify the documents according to their subject themes. Text mining and clustering techniques are common methods used to split a set of data into groups of similar themes (Puteri et al, 2016). Text classification is considered as tagging that classifying text into orderly groups. Text classifiers can analyze a text automatically way, after that, it can determine a set of pre-specify tags or classes based on its content. Unstructured text is difficult to extract value from the data except if it's systematic in a certain way. Text Classification is a task of data mining; it aims to assign automatically selected documents into categories from a pre-defined set of categories (Mahmoud et al, 2017). There are two approaches involved in the processing of text classification: The first approach extracts the feature terms which are recognized as effective keywords in the training phase, and the second is concerned with the actual classification of the document using these feature terms in the test phase, that has been used before in training phase.

The main objective of this paper is to improve the accuracy of Prophetic hadith text by studying the effect of clustering classification and pre-processing of text. Prophetic hadith is what was added to the Prophet Muhammad, it includes the Saying, Doing, Reporting, or Describing of the Prophet. 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, additional to finding a relationship between these classifications. The dataset of hadith in this study 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 (Jiawei et al, 2012). Then the feature selection technique is used to select the most relevant words in each class. After that, the supervised learning algorithms were used to build a classification model capable of classifying the hadith into different categories according to what was attributed to the Prophet mainly: Saying, Doing, Describing, and Reporting. The second part of this study is established to investigate the significant relationship between the types of hadith and its tags using analysis of variance ANOVA and multiple pairwise comparisons. A multivariate technique using Two Way Cluster Analysis was used to classify Hadiths according to its tags. The Experimental Results for Pre-processing of Text were implemented by python 3.6.0 program, Classification using cluster techniques was analyzed using IBM SPSS ver. 23 and PC-ORD ver. 5. The paper is organized as follows: Section I displays a brief background of hadith, clustering techniques and related works. Section II presents hadith classification used in this study. Section III demonstrates the framework for clustering techniques. Finally, the paper end by conclusions and future work in Section 6.

2. Background and Related Work

2.1 Hadith

Hadith is considered as one of the essential Islamic references which considered as the main source of guidance for the Muslims after the Holy Book. Hadith, in the Arabic language, indicates a talk, report, notification, essay, novel or story. Prophetic Hadith is a collection of the says that extract what Prophet Muhammad (PBUH) said and recommended on the Islamic way of life, which agrees with the instruction of Al Quran by trustworthy narrators (Rahman et al, 2010). Also, Hadith can explain by the saying and acts of Prophet Muhammad (PBUH) and reports about his companions which the prophet agreed upon. All these classes are known as the "Sunnah". There are several aspects of human life that covered by Hadith such as Judicial judgments and human relations, commerce, economics political science.

It also deals with family disputes in case of divorce, death, inheritance provisions in cases of death and others which organize social and life relations in accordance with Islamic law (Musa. et al, 2012). After the holy-Quran, Hadith is the most important source of Islamic religious law for Muslims. Each hadith consists of three main parts: the Matn (text of hadith), the Sanad (narrators' chain of hadith: a string of talkers who recording the hadith and reporting it), and the Taraf (the part, or the beginning sentence of the text that refers to the sayings, characteristics or actions of the Prophet, or his concurrence with others action) (Ismail, et al, 2014). The first Muslim scientists classified the Hadith also, based on the degree of uniqueness and accuracy into four categories. The categories of Hadith are (i) "Sahih" refers to correct, true, valid. (ii) "Hasan" related to its prophet properly, but the narrators found in it some weakness; (iii)

"Da'if" is basically a weak hadith which makes it unreliable and acceptable; (iv) "Maudu" which is attributed to Prophet Muhammad (PBUH) peace be upon him - and he did not say or do or acknowledge it. The two most highly respected collections of Hadith are Sahih Bukhari, and Sahih Muslim follow by the other four collections are the Sunan of Tirmidhi, Nasa'i, Ibn Majah and Abu Da'ud. These four collections and two Sahih collections were formed as a "six books" or Al-Kutub al-Sitta (Puteri et al, 2016).

2.2 Clustering

Clustering is one of the data mining techniques that have been widely applied in text mining. Cluster technique is considered as an unsupervised learning algorithm whose goal is to classify data into similar groups that have the same characteristics (Akbar, 2008), (Ding and Fu, 2012). Clustering has wide applications in the context of text especially in classification, visualization and document organization. Text clustering can be in various scales of similarity, it could be documents, passages, sentences or expressions (Allahyari et al, 2017). The main Text clustering algorithms are divided into many different types such as agglomerative clustering, partitioning, and probabilistic clustering algorithms. Many studies have been conducted to explore the implementation of text clustering algorithms in text documents.

The most common text clustering algorithms are Hierarchical Clustering algorithms, k-means Clustering, and Probabilistic Clustering and Topic Models. K-means and K-medoids clustering are widely used in partition algorithms (Berkhin, 2017).

The hierarchical clustering algorithms can be categorized into agglomerative (bottom-up) and divisive (top-down). Hierarchical

clustering algorithms are one of the Distanced-based clustering algorithms. Hierarchical clustering is often displayed graphically using a tree diagram called a dendogram which displayed cluster and sub-cluster relationships (Allahyari et al, 2017).

In this study, two-way cluster analysis was used to classify Hadiths according to the tags present in each Hadith. Linkage Clustering and two-way dendogram was created to show clusters of Hadiths and clusters of tags. A Linkage Clustering is dependent first on lessening similarities or grow distance with the most identical cases and continuing until all pairs are considered for. Second, the clusters are created hierarchically, beginning with the most identical cases, and then keeping the cases cluster into groups. A dendrogram is the most commonly used method of summarizing the hierarchical clustering results. In hierarchical clustering, it clarifies the order of the clusters generated by the corresponding analyses.

2.3 Related Work

There are several types of research have been studied to solve the problem of Arabic document classification, while a few of researches have been studied in hadith text. In this section; we show some of the related works in the hadith classification as shown in table 1.

Table 1:" some of the related works in the hadith classification".

Published by	Output	Method/Technique
Aldhlan et al, (2013)	Authentication into Sahih, Hasan, Da'if and Maudu	Appending Isnad (manually), Eliminating punctuation and diacritical signs (manually), Append specific character (manually), Classify Hadith (Decision Tree optimized using Missing Data Detector)
Halim et al, (2007)	Authentication status of Hadith used in Tafsir Al- Azhar.	Manually Takhreej Hadith
Dad et al, (2014)	Authentication Status of Mursal Hadith	Manually Literature Analysis on Hadith Sanad
Ibrahim et al, (2017)	Hadith Authentication into Sahih or not Sahih	Takhreej Al-Isnad combined with a new mechanism which inspired from a prophetic strategy the battle of Badr
Kamsin et al, (2015)	Hadith Authentication into Sahih, Hasan, Da'if	Takhreej Hadith using a Unicode centric string-matching approach
Aldhaln et al, (2012)	Hadith Authentication into Sahih, Hasan, Da'if and Maudu	Takhreej Hadith combined with Decision Tree (DT) classifier and Missing Data Detector (MDD

Ghanem et al, (2016)	Hadith Authentication into Sahih, Hasan, Da'if and Maudu'	Removing Matn (manually), Removing verbs (manually), Names Standardization (manually), Feature extraction (TF-IDF), Document representation (VSM), Classification (LVQ)
Ghazizadeh et al, (2008)	Narrator Reliability and Hadith Validity Level	Science of Hadith Rijal al-Hadith combined with Fuzzy Rule-Based System
Rozi et al, (2008)	Status of Hadith authentication into Maqbûl and Mardûd	Sanad connection rate, Jarh wa Ta'dil Level, Hadith Classification Modeling using Naive Bayes (NB), Hadith Searching Modeling using Vector Space Model (SVM)

3. Hadith Classification

The two essential parts of Hadith are the Sanad and the Matn (Text). In this classification, the hadith has been classified on the Hadith text only, and the Sanad is neglected, where it is irrelevant to hadith categories. The hadith has been classified into four categories mainly:

Describing, Saying, Reporting and Doing according to what was attributed to the Prophet Muhammad (PBUH). Hadith text has been processed such as other Arabic texts rest to convert it to a form that is convenient for classification tasks, through the following steps.

Hadith Collection

In this study, a number of hadiths were collected from the highest reference collections of Hadith such as Sahih Bukhari, and Sahih Muslim, to study the effect of the text pre-processing on hadith classification, according to what was attributed to the Prophet (PBUH). Hadiths collected for each category were shown in Figure 1, which reached up to 1458 hadiths followed the pre-processing technique.

Hadith Pre-Processing

Data pre-processing is a process of applying various techniques over the raw data through a series of improving quality techniques to make it ready for processing. Pre-processing is an important technique to decrease the vagueness of the words and increase the classification accuracy. It includes the normalization, tokenization, stop word removal and stemming (Abdullah et al, 2016; Mubarak et al, 2015; Alajmi and Said, 2012; and Zhengwei et al, 2010). Figure 2 shows the main steps of pre-processing by an example of hadith.

Parts of Speech

Parts of speech are the process of assigning each word to its role in the sentence, named tags. These tags represent grammatical classification based on eight parts of speech: the verb (VB), the noun (NN), the pronoun (PR+DT), the adjective (JJ), the adverb (RB), the preposition (IN), the conjunction (CC), and the interjection (UH). It also may include additional information with status tags like a number, gender, etc. There are two language processing systems that usually used to follow the parts of speech recognition tools. The oldest one which uses a dictionary for reference, and the random method that uses probabilistic and statistical information to assign the tag to a word (Mubarak, 2015).

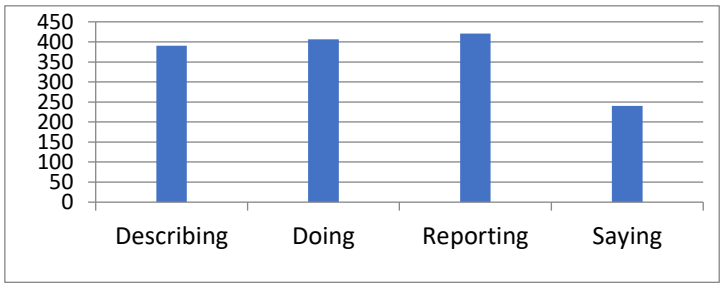

Figure 1 Number of collected Hadith per Category

Document Representation

One of the pre-processing techniques is document representation. This technique is used to represent the documents as a vector or for converting full text into vectors such as index terms, which most used in information retrieval and text classification. There are several methods used to determine the weight for each word in the document, one of the most common methods is TF-IDF. It is a statistical method used to determine the frequency of each word in a hadith text according to Equation 3. TF-IDF method aims to reflect the importance of the word to the text in a dataset, which helps us to determine the important words in a document collection for classification purposes (Abdelaal, 2019).

$$TF_{i,j} = \frac{n_{i,j}}{\sum n_{k,j}} = \frac{f_{w,t}}{\max \{f_{w',t} : w' \varepsilon t\}} \quad (1)$$

Where: $n_{i,j}$: is the number of occurrences of the word (w_i) in the text (t_j). $\sum n_{k,j}$: is the sum of the number of occurrences of all terms in a text.

$$IDF\ (w,\ t) = log \frac{N}{n} = log\ (\frac{N}{|\{t\ \varepsilon\ D : w\ \varepsilon t\}|}) \quad (2)$$

Where: $IDF_{(w,\,t)}$ is to calculate the Inverse Document Frequency, N is the total number of texts t in the dataset of hadith n is the number of texts t that contains a word w.

$$TD\text{-}IDF = TF_{i,j} * IDF\ (w,t) = \frac{n_{i,j}}{\sum n_{k,j}} x\ log\ \frac{N}{|\{t\ \varepsilon\ D{:}w\ \varepsilon t\}|} \qquad (3)$$

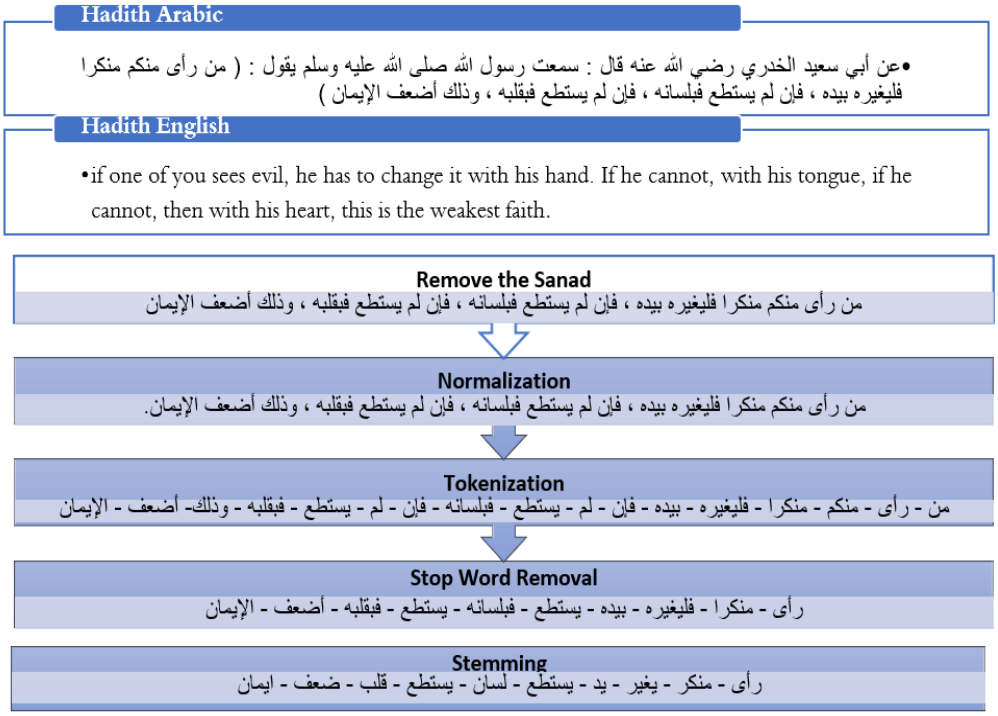

Figure 2 Main steps of pre-processing in text classification

Feature Selection

Feature selection (FS) is the next step after pre-processing and text representation using term weighting. Feature selection helps to identify the related features from a set of data and removing the irrelevant or less important features. Reducing the number of features leads to improve accuracy, reducing risks and improve visualizing data. In addition, it

enables the learning algorithm to train faster (by Forman. 2003 and Hassan.1996). There are several techniques used to select the best features, as probability ratio, Gini Index, Principal Component Analysis, Latent semantic analysis, Gain Ratio, Symmetrical uncertainty, odds ratio, with taking into account the constraints in the selection of feature (by. Bassinet et. Al. 2018).

Classifier Learning

The essential goal of Classifier learning is to train the classifier to distinguish the main features of the training dataset for each category and predict the category of hadith. The general approaches for building a classification model is shown in Figure 3

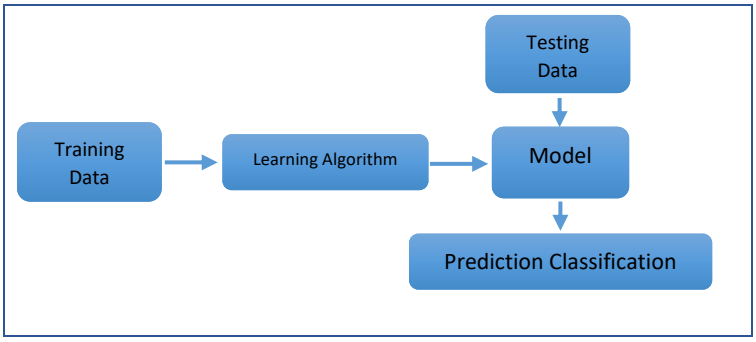

Figure 3 approaches of classification model

In this study, the cross-validation technique was used by dividing the dataset into two groups: training and testing. The dataset contains 1458 hadith, were divided randomly into a number (n) of blocks. Each block is held out once to test the classifier, and the classifier is trained on the remaining (n-1) to build the classifier (Han et al, 2012). Different algorithms were used to construct a model able to classify the class of hadith.

Experimental Results

In this classification, the dataset of hadith was divided randomly into 10 parts, each part of them was held out once to test the classifier, and the classifier was trained on the remaining nine parts to train and test all dataset known as cross-validation method (Han et al, 2006). The results are summarized in Table 2, and Figure 4 shows the overall accuracy for each classifier. These Experimental Results were implemented by python 3.6.0 program. Different algorithms have used in this study, but only the best three algorithms were reported, which satisfied the highest accuracy.

Table 2: Overall Accuracy for each Classifier

Classifier	SGD Classifier	Logistic Regression	Bernoulli NB
Accuracy	٠,٩٦٨٤٠٧٨	٠,٩٥٦٠٦٩٠	0.93829

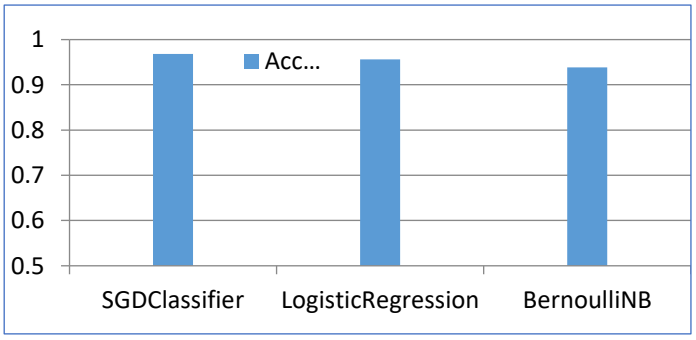

Figure 4 percentage accuracies for individual classifier

Table 3 and Figure 5 show the Precision, Recall, and F1-measure (F1) for each category using SGD Classifier, Logistic Regression, and Bernoulli NB. These categories are Describing, Doing, Reporting, and Saying, according to what was attributed to the Prophet using the cross-validation

method. We note from the values of Precision, Recall, and F1 there no bias between these values, which indicates the efficiency of the classifier is more general.

Table 3: Precision, Recall, and F1-score for each category

Classifier	SGD Classifier			Logistic Regression			Bernoulli NB		
Class	Precision	Recall	F1-measure	Precision	Recall	F1-measure	Precision	Recall	F1-measure
Describing	0.93	0.98	0.95	0.90	0.97	0.93	0.96	0.87	0.91
Doing	0.97	0.99	0.98	0.97	0.99	0.98	0.91	0.98	0.94
Reporting	0.99	0.92	0.95	0.98	0.90	0.94	0.92	0.94	0.93
Saying	0.99	1	0.99	1	0.97	0.99	0.99	0.98	0.99
Avg	0.97	0.97	0.97	0.96	0.96	0.96	0.94	0.94	0.94

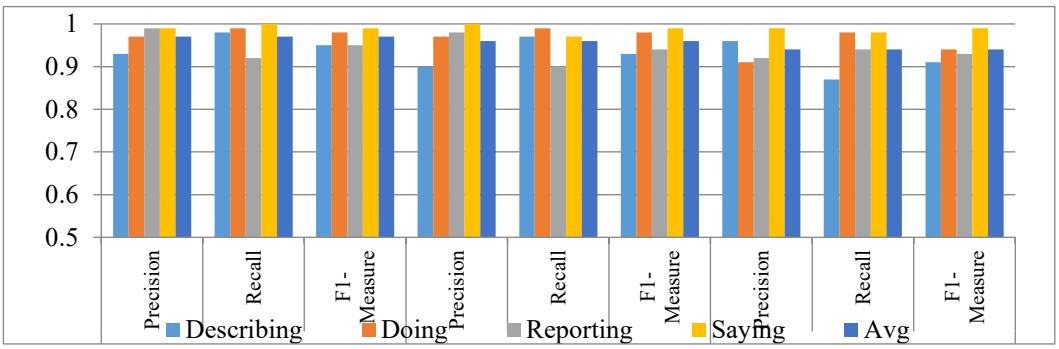

Figure 5: Precision, Recall, and F1-measure for each category

The Framework For Clustering Technique

This section investigates the presence of a relationship between the classify hadiths, and the significant association between the degree of the truth of hadith if Sahih, Hasan, Da'if, or Maudu and its type if Saying, Reporting, Doing, or Describing. A set of 400 hadith was used to study if

tags could be used to classify different sorts of Hadith using ANOVA and multiple comparisons methods. Finally, Two Way Cluster Analysis was used to classify Hadiths according to the frequency of the tag in each Hadith.

Relation between Classifications

The number of the Prophet's Hadiths cannot be precisely confirmed or restricted at all. There is no definite reference that can determine a fixed number if hadith is Sahih, Hasan, Da'if, or Maudu. If we want to count the number of valid hadith only, we will not be able to do so, as this will vary from one reference to another. By referring to the documented six books of Hadith (*six major hadith collections in Sunni Islam): (Bukhari, Muslim, Sunan al-Tirmidhi, Nesa'y, Abu Dawood, and Ibn Majah) (by Al-Bukhari. 9th century). In this study, a summary collection of classified hadiths collected in in Table 4. The hadith were classified into two classifier groups: group 1 refer to (Sahih, Hassan, Da'if, and Maudu) and group 2 refer to (Saying, Reporting, Doing, and Describing). It found that there are around 5109 hadith classify for group 1 and 6857 for group 2.

Ṣaḥīḥ al-Bukhārī is one of the Kutub al-Sittah (six major hadith collections) of Sunni Islam. Whereas, out of all these six major books, the collection of prophetic traditions, or hadith for Sahih al-Bukhari, was performed by the Muslim scholar Muhammad al-Bukhari. It was completed around 846 AD / 232 AH. Sunni Muslims view this as one of the two most trusted collections of hadith along with Sahih Muslim.

Table 4: Summary Collection of Classification Hadiths

Group 1	%	Group 2	%
Sahih	1384 (27.1%)	Saying	1741 (25.4%)
Hassan	1394 (27.3%)	Reporting	1625 (23.7%)
Da'if	1294 (25.3%)	Doing	1732 (25.3%)
Maudu	1037 (20.1%)	Describing	1759 (25.6%)
Total	5109 (100%)	Total	6857 (100%)

In this section, we want to detect if there is a relationship between the classification of the degree of credibility of the hadith (Sahih, Hasan, Da'if, and Maudu) and the classification in terms of (saying, doing, describing, or reporting). A random sample of 2929 hadith was selected and summarized in Table 5. The results found that there are around 28% (819 hadith) classify as Sahih, 30% of them are classifying as "doing". And 14.3% (419 hadith) classify as Maudu, and 29.3% classify as Da'if that the percentage of weak hadith is near to 43.6%. Chi-square test was used to test if there is an association exists between the type of hadih if it (Sahih, Hassan, Da'if, and Maudu; group 1) and its kind if it (Saying, Reporting, Doing, and Describing; group 2). We found that chi-square= 214.6 with ($p < 0.001$); this mean that the hypotheses that there is a strong relationship between groups is accepted. This concludes that group1 seems to be significantly related to group 2 with 5% significant level.

Table 5: Cross-Tabulation Chi-Square Test

	Sahih	Hasan	Da'if	Maudu	Total
Saying	131(16%)	165(20%)	312(36.4%)	90(21.5%)	698(23.8%)
Doing	248(30%)	258(31%)	113(13.2%)	133(31.7%)	752(25.7%)
Describing	203(25%)	166(20%)	278(32.4%)	91(21.7%)	738(25.2%)
Reporting	237(29%)	244(29%)	155(18%)	105(25.1%)	741(25.2%)
Total	819(28%)	833(28.4%)	858(29.3%)	419(14.3%)	2929(100%)

Experimental Results and Analysis

A sample of 400 hadith was selected, 29 tags are founded and classified for terms of hadith as Sahih, Hasan, Da'if, and Maudu. Data were analyzed using IBM SPSS ver. 23 and PC-ORD ver. 5. The custom table has been generated to show frequencies sum and mean of each tag within each group of Hadith. Results summarized in Table 6 showed that the frequency of replicated tags can be classified into two similar groups, which are "Sahih & Hassan" and "Da'if & Maudu".

Analysis of variance test (ANOVA) was used to compare groups of Hadith from the perspective of both degree of correctness and Hadith type for each tag as univariate analysis. Multiple pairwise comparisons among groups were established using Holm-Sidak method, and custom tables for significant results were summarized in Tables (7) and (8) respectively. Results showed that some of the tags could distinguish the difference between types of hadith. The most results found in a custom table for significant Tags agreed with the hypotheses that there are significant differences between Da'if and (Hassan & Sahih) but not always differences between Da'if & Maudu.

Table 6: Frequencies Sum & Mean of Tags

Tags	Da'if		Hasan		Sahih		Maudu	
	Sum	Mean	Sum	Mean	Sum	Mean	Sum	Mean
JJ	56	.6	62	.6	63	.6	65	.7
NN	745	7.5	894	8.9	878	8.8	710	7.1
NNS	12	.1	31	.3	25	.3	33	.3
NNP	687	6.9	544	5.4	562	5.6	341	3.4
PRP	8	.1	1	.0	5	.1	11	.1
RP	74	.7	42	.4	50	.5	46	.5
VB	1	.0	8	.1	4	.0	7	.1
VBD	196	2.0	295	3.0	354	3.5	206	2.1
VBG	8	.1	0	0.0	2	.0	1	.0
VBN	14	.1	6	.1	7	.1	11	.1
VBP	154	1.5	136	1.4	155	1.6	131	1.3
WP	33	.3	16	.2	32	.3	20	.2
DTNN	155	1.6	299	3.0	257	2.6	169	1.7
IN	137	1.4	182	1.8	186	1.9	149	1.5
CC	13	.1	28	.3	23	.2	13	.1
CD	10	.1	18	.2	8	.1	3	.0
DT	11	.1	33	.3	17	.2	16	.2
DTJJ	5	.1	15	.2	8	.1	18	.2
DTNNP	21	.2	36	.4	25	.3	9	.1
DTNNS	5	.1	0	.1	11	.1	12	.1
NOUN	4	.0	6	.1	13	.1	3	.0
WP	33	.3	16	.2	32	.3	20	.2
DTJJR	0	0.0	0	0.0	0	0.0	4	.0
UH	1	.0	0	0.0	1	.0	0	0.0
ADJ	0	0.0	1	.0	0	0.0	0	0.0
JJR	1	.0	1	.0	0	0.0	3	.0
PRP$	0	0.0	2	.0	1	.0	0	0.0
VN	1	.0	0	0.0	1	.0	2	.0
RB	1	.0	1	.0	1	.0	0	0.0

Table 7: Results of ANOVA for significant Tags

Tags	F	Sig
NN	11.492	.000
NNP	27.615	.000
RP	3.630	.013
VBD	23.804	.000
VBG	4.961	.002
DTNNP	5.203	.002
IN	4.328	.005
DT	4.420	.005
DTJJ	2.787	.040
NOUN	2.907	.035

Table 8: Multiple Pairwise Comparison among Groups

Tags		Sig		
		Sahih	Hassan	Da'if
NN	Hassan	.999	-	-
	Da'if	.004	.001	-
	Maudu	.000	.000	.936
NNP	Hassan	.998	-	-
	Da'if	.008	.001	-
	Maudu	.000	.000	.067
	Hassan	.973	-	-
	Da'if	.041	.017	-

RP	Maudu	.029	.047	.053
	Hassan	.053	-	-
	Da'if	.000	.000	-
VBD	Maudu	.000	.000	.998
	Hassan	.944	-	-
VBG	Da'if	.052	.003	-
	Maudu	.048	.049	.064
	Hassan	.170	-	-
DTNN	Da'if	.000	.000	-
	Maudu	.000	.000	.978
	Hassan	.999	-	-
IN	Da'if	.019	.039	-
	Maudu	.142	.245	.977
	Hassan	.076	-	-
DT	Da'if	.025	.004	-
	Maudu	.039	.049	.968
	Hassan	.999	-	-
DTJJ	Da'if	.046	.049	-
	Maudu	.025	.025	.052
	Hassan	.319	-	-
NOUN	Da'if	.045	.055	-
	Maudu	.046	.963	0.999

As a multivariate technique, Two Way Cluster Analysis was used to classify Hadiths according to the tags present in each Hadith. Euclidean distance was used as a distance measure and Ward's method was used as a linkage method. The two-way dendrogram was created, Figure 6 shows clusters of Hadiths and clusters of tags and a heat map that represents the intensity of each tag in each Hadith. It consists of two main groups: two similar means were between Sahih and Hassan as category and the remainder classes together similar are Da'if and Maudu which indicator to an expected significant difference between tags. the common tags in this way show that there are two groups first we test the significant difference and then multiple to know where is the different exacted then use a cluster to classify.

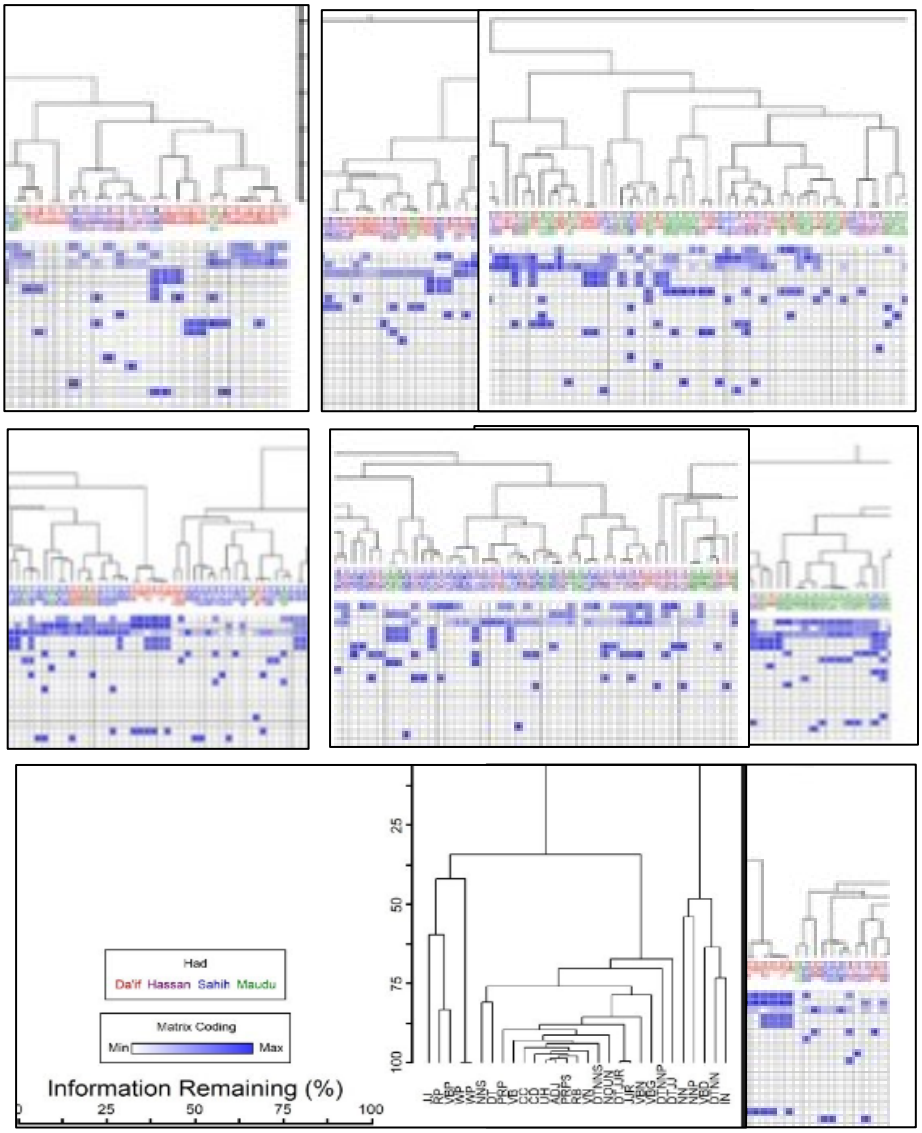

Figure 6: The Two-Way Dendrogram

Conclusion

The importance of studying hadith is due to that it is the main source of Islamic religious law for Muslims after the holy-Quran, so considered as the second source of Islam and the fundamental resource of legislation in the Islamic community. The results in this study showed that there is an effect of text pre-processing on hadith text classification according to what was issued from Prophet Muhammad (PBUH). It found that the best classifier had given high accuracy was SGD Classifier; it achieved higher accuracy reached up to 0.9684 % followed by Logistic Regression, reached up to 0.9560, and finally, Bernoulli NB reached up to 0.9383. Results also concluded that there is a significant association between the type of hadith (Sahih, Hassan, Da'if, and Maudu) and the classification (Saying, Reporting, Doing, and Describing).

The experimental results in this study found that there is a relation between the tags of hadith text and its class according to what was Prophet Muhammad (PBUH) said or did. This relation helps the classifier is able to realize and characterize the hadith classes. Univariate analysis using ANOVA and pairwise comparisons showed that Tags can distinguish that there are significant differences between Da'if and (Hassan & Sahih) and non-significant differences between Da'if & Maudu. The multivariate technique, using Two Way Cluster Analysis results showed that a number of tags could be used to segregate different sorts of Hadith especially the Hadith which classify as Da'if and Maudu as the classification results showed distinct groups of both types while classification showed low accuracy or capability of the separation between Sahih and Hassan Hadiths. Applying the experiment with the largest number of Hadiths may get different results to confirm or disagree with the obtained in this study. So, as future work we suggest additional studies and analysis working with more numbers of hadith which will be supporting this classification and will be more determine the groups of Sahih hadiths. This will be very important because it will resolve conflicting views on hadiths which classify as Da'if and Maudu.

References

1. Abdelaal, H. M., Elemary, B. R. and Youness. H. A. (2019). Efficient Feature Representation Based on the Effect of Words Frequency for Arabic Documents Classification. IEEE Access, 7, 152379 – 152387.
2. Abdullah, A., Guanzheng TAN, Khaled A, and Rajeh, H. (2016). The Effect of Pre-processing on Arabic Document Categorization. Algorithms.
3. Akbar, M. (2008). FP-growth approach for document clustering. Montana State University-Bozeman, College of Engineering.
4. Alajmi, A. and Said, E. M. (2012). Toward an ARABIC Stop-Words List Generation. International Journal of Computer Applications, 46(8).
5. Aldhaln, K.A., Zeki, A. M. and Zeki, A. M. (2012). Knowledge extraction in Hadith using data mining technique. Proc. 2nd Int. Conf. E Learning Knowl. Manag. Technol. (ICEKMT 2012), 2, 13–21.
6. Aldhlan, K. A., Zeki, Ak. M., Zeki Ah. M., and Alreshidi, H. A. (2013). Novel mechanism to improve hadith classifier performance, in Proceedings-2012 International Conference on Advanced Computer Science Applications and Technologies, 512–517.
7. Allahyari, M. Pouriyeh, S, Assefi; M, Safaei; S, Trippe; E, Gutierrez, J and Kochut; K. (2017). A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. KDD Bigdas, Halifax, Canada.

8. Bassinet, S., Madani, A., Al-Sarem, M..and Kissi, M. (2018). Feature selection using an improved Chi-square for Arabic text classification ‘, Journal of King Saud University, Computer and Information Sciences’, Elsevier.
9. Berkhin, P. (2006). A survey of clustering data mining techniques grouping multidimensional data. Springer, 25-71.
10. Mubarak, D. S. Madhu, and Shanavas. S A. (2015). A New Approach to Parts of Speech Tagging in Malayalam. International Journal of Computer Science & Information Technology (IJCSIT), 7(5).
11. Dad, K. and Shafiq, M. S. (2014). Mursal Hadith & its Authenticity: A critical analysis. Acta Islam. 2(1), 21–33.
12. Ding, Y., and Fu, X. (2012). Topical Concept Based Text Clustering Method. Paper presented at the Advanced Materials Research.
13. Forman, G. An. (2003). The extensive empirical study of feature selection metrics for text classification. J. Mach. Learn. Res. 3, 1289–1305.
14. Ghanem, M., Mouloudi, A. and Mourchid, M. (2016). Classification of Hadiths using LVQ based on VSM Considering Words Order.' Int. J. Comput. Appl., 148(4), 25–28.
15. Ghazizadeh, M. Zahedi, M. H, Kahani, M. and Bidgoli B. M. (2008). Fuzzy Expert System in Determining Hadith1 Validity. In Advances in Computer and Information Sciences and Engineering, Dordrecht: Springer Netherlands, 354–359.
16. Halim, ‘M., Hamka Dan Tafsir Al-Azhar. (2007). Suatu Kajian Kualiti Hadith. University Saints Malaysia.

17. Han, J., Kamber, M. and Pei, J. (2006). Data Mining: Concepts and Techniques", the Second Edition University of Illinois at Urbana Champaign.
18. Han, J., Kamber, M., and Pei, J. (2012). Data Mining: Concepts and Techniques. The Third Edition University of Illinois at Urbana Champaign, Elsevier Inc. Web site at www.mkp.com.
19. Hassan, S. (1996). Introduction to the Science of Hadith Classification. Darussalm, Riyad, Saudi Arabia.
20. Ibrahim, N. K., Samsuri, S. Seman, M. S. A. Ali, A. E. M. B. and Kartiwi, M. (2017). Frameworks for a Computational Isnad Authentication and Mechanism Development. In Proceedings - 6th International Conference on Information and Communication Technology for the Muslim World, ICT4M, 154–159.
21. Kamsin, A. (2015). Program for Developing the Novel Quran and Hadith Authentication System. International Journal on Islamic Applications in Computer Science and Technology.
22. Mahmoud, A., Khan, H., Rehman1, Z., and Khan, W. (2017). Query-based information retrieval and knowledge extraction using Hadith datasets. Conference Paper. https://www.researchgate.net/publication/323193943.
23. Muhammad al-Bukhari. (9th century). Sahih al-Bukhari. Kutub al-Sittah Hadith, 846 AD, Dar Al-Nawader, (Muslim ibn al-Hajjaj, "Ṣaḥīḥ Muslim;" Kutub al-Sittah, Dar Al-Khalafa Al-aliya,).
24. Musa, S. A., Ahmed, A. F., & Mustapha, A. R. (2012) Studies on the Hadith: National Open University of Nigeria.

25. Nanang Fakhrur Rozi, E. M. K, and N. R. M. (2008). Identifikasi Jenis Hadits Menggunakan Beberapa Kombinasi Metode Learning.
26. Puteri, N. E., Nohuddin1, A., Zuraini Zainol, B., Kuan Fook Chao, C, A. Imran Nordin1, D, and Tarhamizwan M. James, A. H. (2016). Keyword based Clustering Technique for Collections of Hadith Chapters. International Journal on Islamic Applications in Computer Science and Technology, 4(3), 11-18.
27. Qiu, Z. Gurrin, C., Doherty, A. R. and Smeaton, A. F. (2010). Term Weighting Approaches for Mining Significant Locations from Personal Location Logs. 10th IEEE International Conference on Computer and Information Technology, Bradford, 20-25. doi: 10.1109/CIT.2010.48
28. Rahman, N. A., Bakar, Z. A., & Sembok, T. M. T. (2010). Query expansion using thesaurus in improving Malay Hadith retrieval system. Paper presented at the Information Technology (itsim), International Symposium in.
29. Tuan Ismail, R. Baru2, A. Hassan, and Salleh, A. (2014). The Matan and Sanad Criticisms in Evaluating the Hadith. Asian Social Science; Published by Canadian Center of Science and Education, 10(21).