# The Impact of Text Pre-processing and Term Weighting on Al-Hadith Al-Shareef Classification

Ahmed S. J. Abu Hammad

***Abstract*—**Preprocessing is one of the key components in a typical text classification framework. The preprocessing step usually consists of tasks such as tokenization, filtering, lemmatization and stemming. This paper studies the impact of text pre-processing and totally different term weighting schemes on Al-Hadith Al-Shareef Classification. Additionally, thereto, presents and compares the effectiveness of three distinct automatics learning algorithms for classifying Al-Hadith Al-Shareef into eight selective books depending on Sahih Muslim. To the best of our knowledge, there is still no published study on this data set. The automatic learning algorithms are Naïve Bayes (NB), Support Vector Machines (SVM), and Complement Naïve Bayes (CNB) with 10-fold cross-validation. We used Term Frequency-Inverse Document Frequency (TF-IDF), Term Frequency (TF), Term Occurrences (TO), and Binary Term Occurrences (BTO) techniques to compute the relative frequency for every word in a very specific document. The results indicate that term stemming and pruning, document normalization, and term weighting dramatically reduce reductional, enhance text representation and directly impact text mining performance. What is more, classification results show that the CNB achieved promising results compared with other supervised methods in classifying A-Hadith. CNB obtains 91.22% accuracy and 91.86% F-measure.

***Keywords*—** Arabic Text Classification, Arabic Text Mining, Arabic Morphological Analysis, Term weighting.

## I. INTRODUCTION

Islam based on two fundamental laws: Al-Qur'an as the set of words of Allah and Al-Hadith that documenting words, deeds, provisions, and approvals of Mohammad as the prophet of Allah. Hadith was compiled and classified by many Imams such as Imam Bukhari, Imam Muslim, and Imam Tirmidzi, etc. All of them based on one source prophet Mohammad (peace and blessings of Allah be upon him). Imam Muslim is one amongst the known Imam that according to Ulama. Imam Muslim spent nearly fifteen years to compile over 3000 Hadiths without repetition [25]. Referring to [28], Figure 1 is the component of Hadith.

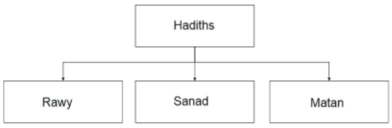

Fig. 1 Hadith components.

Sanad is that the chain of the conveyor of every Hadith, this part present at the beginning of Hadith. Matan is that the content of Hadith, present after the Sanad, and at last Rawy, this is the person or Imam that compile Hadith such as Imam Muslim.

By the exponential growth of digitalized document, emerge the necessity of a system that ready to extract high-quality information, that's why automatic Text Classification (TC) become widespread.

TC task goes through three main steps: text pre-processing, text classification and evaluation. Text pre-processing phase is to make the text documents appropriate to train the classifier. Then, the classifier is built and tuned employing a learning technique against the training

Ahmed Abu Hammad, University College of Science and Technology, Khan Younis, Palestine (e-mail: asj_hammad@hotmail.com).

dataset. Finally, the classifier gets evaluated by some evaluation measures, i.e. recall, precision; etc. The careful description of those steps is often found in [29, 30, 31].

Several existing classification algorithms are used to classify English text corpora such as: SVM [6, 33], NB [6, 7, 33], NB [6, 7, 33], Decision Trees (DTs) [6, 7], k-Nearest Neighbor (KNN) [33], Artificial Neural Networks (ANNs) [33] et al. However, little research works are conducted on Arabic corpora, chiefly since the Arabic language is very wealthy and needs special treatments like order verbs, morphological analysis, etc. Notably, in Arabic morphology, words have affluent meanings and contain a good deal of grammatical and lexical information [32]. Additionally, in grammar structure, Arabic sentence formation differs from English. During this regard, the Arabic text documents are required, significant processing to build an accurate classification model. Therefore, few scholars have applied a variety of classification approaches to the matter of Arabic text classification, i.e. NB [3, 10] [13], SVM [2, 15, 22], KNN [22] and DTs [2, 16]. Even so, researchers conclude that the Arabic text classification may be a terribly difficult task because of language complexity.

This paper studies the impact of text pre-processing techniques and different term weighting schemes on Arabic corpus collected manually from Islam's lawsuit and indicative website. Additionally, presents and compares varied classification rules mining methods associated with the matter of Arabic text classification. Primaries, NB, SVM, and CNB learning methods are applied to classify Sahih Muslim Arabic corpus into one of the predefined categories (books) to measure their performance and effectiveness with reference to different text evaluation metrics like accuracy, precision, recall, and F-measure measures. Experiments are going to be conducted on a specific set of AL-Hadith from Muslim book, wherever eight selective books were chosen as categories so as to run these experiments.

The sub-sequence sections are organized as follows: section 2 contains related works. Section 3 introduces the corpus; we used to test our learning methods and the pre-processing done to the text. Finally, experimental results and evaluation, and conclusions are presented in Section 4 and Section 5 respectively.

## II. Related Works

The Arabic language is the mother tongue of more than 300 million people; it is considered for religious reasons the language of Islam, and it is ranked as the fifth most spoken language around the world [26]. Unlike Latin-based alphabets, the orientation of writing in Arabic is from right to left; the Arabic alphabet consists of 28 letters. Arabic, in general, is a challenging language because it has a very complex morphology as compared to English. This is due to the unique nature of Arabic morphological principle, which is highly inflectional and derivational [9, 11, 14].

El-kourdi et. al [10] used an NB classifier to classify an in-house collection of Arabic documents. The collections include five classes and three hundred web documents for every class and have used many partitions of the data set. They have concluded that there is some indication that the performance of the NB algorithm in classifying Arabic documents is not sensitive to the Arabic root extraction algorithm, additionally to their own root extraction algorithm; they used other root extraction algorithms. The average accuracy reported was about 68.78%.

Duwairi [8] compared the performance of NB, KNN, and distance-based classifiers for Arabic text categorization. The collected corpus contains a thousand documents that vary in length and writing styles and comprise ten classes every class consists of a hundred documents. The author used stemming to reduce the number of features extracted from documents. The recall, precision, error rate and fallout measures were used to compare the accuracy of classifiers. The results showed that the performance of NB classifier outperformed the other two classifiers.

Al-harbi et. al [2] evaluated the performance of two popular classification algorithms C5.0 decision tree and SVM on classifying Arabic text using the seven different Arabic corpora such as (Saudi News Papers, WEB Sites, Arabic Poems). They have implemented a tool for Arabic text classification to accomplish feature extraction and selection. They have concluded that the C5.0 decision tree algorithm outperformed SVM in terms of accuracy whereas the SVM, average accuracy was 68.65%, while the average accuracy for the C5.0 was 78.42%.

Hattab et. al [17] applied the SVM model in classifying Arabic text documents. The results compared with the other traditional classifiers NB classifier, KNN classifier, and Rocchio classifier. Their experimental results performed on a set of 1132 documents, showing that Rocchio classifier gave better results when the size of the feature set is small while SVM outperformed the other classifiers when the size of the feature set was large enough. The classification rate exceeds 90% when using more than 4000 features.

Al-khatib [4] compared the effectiveness of four different learning algorithms for classifying Al-Hadith Al-Shareef into eight selective books depending on Sahih Bukhari. The testing corpus has 1500 Hadiths that vary in length distributed eight books. The learning algorithms are the Rocchio algorithm, KNN, NB, and SVM. He used the Term TF-IDF technique to compute the relative frequency for each word in a particular document. His results showed that the best accuracy was reported for the SVM algorithm in AL-Hadith Classifications since the precision value is the smallest one for all results. KNN and NB algorithms had a good accuracy in Al-Hadith classifications, and the worst accuracy is reported for the Rocchio algorithm in AL-Hadith classifications since the precision value is the largest one.

Jbara [19] examined the knowledge discovery from AL-Hadith through a classification algorithm in order to classify AL-Hadith to one of the thirteen predefined classes (books) from Sahih AL-Bukhari. The testing corpus has 1321 Hadiths that vary in length distributed over thirteen books. The author used a supervised method called Stem Expansion (SEC) to discover knowledge from AL-Hadith by assigning each Hadith to one book (class) of predefined classes. His results showed that SEC performed better in classifying AL-Hadith against existing classification methods (WBC and AL-Kabi) according to the most reliable measurements (recall, precision, and F-Measure) in the text classification field.

We found that there's a significant shortcoming of the Arabic classification studies during this field. Each study is restricted to a limited range of classification algorithms. This research studies the impact of text pre-processing and different term weighting schemes on Arabic text classification. Additionally, presents and compares distinct classification methods that may use the same corpus in order to evaluate such algorithms and choose the one most suited to the considered case study. This guarantees that the various algorithms had the same conditions and also the same setting in all the experiments.

## III. The Corpus and The Text Pre-processing

### *A. The Corpus*

In this work, we tend to build an in-house corpus of Arabic texts collected from [18], that referred to as MHAC to perform our experimentation; the corpus includes 1,306 text documents and classified in eight classes that chosen from Sahih Muslim. The corpus contains concerning 24,127 district features after stop words removal. We generate all text representations for MHAC corpus to assess the obtained classification results. The generated text representations for MHAC corpus are: (Light stemming, Stemming) and percentual term pruning (min threshold = 3%, max threshold = 30%) with (TF-IDF, TF, TO, and BTO). Table 1 shows statistical information concerning the books included within the experiments along with its name in English and Arabic as it was used by Sahih Muslim.

TABLE I
UNITS FOR MAGNETIC PROPERTIES

| Book (Class) Name | اسم الكتاب | Number of text documents | Number of distinct features after stop words removal |
|---|---|---|---|
| The Book of Prayers | كتاب الصلاة | 238 | 4062 |
| The Book of Zakat | كتاب الزكاة | 168 | 4758 |
| The Book of Fasting | كتاب الصيام | 200 | 3050 |
| The Book of Marriage | كتاب النكاح | 124 | 2602 |
| The Book of Transactions | كتاب البيوع | 115 | 1021 |
| The Book of Musaqah | كتاب المساقاة | 131 | 2266 |
| The Book of Drinks | كتاب الأشربة | 185 | 3454 |
| The Book of Greetings | كتاب السلام | 145 | 2914 |
|  | Total | 1306 | 24127 |

### B. The Text Pre-processing

One of the widely utilized methods for text mining presentations is viewing the text as a Bag of Tokens (BOT) (words, n-grams). Under that model, we can already classify text [5].

Before applying any algorithm, for both training and testing data, some pre-processing will be conducted on each Hadith. It includes removing Sanad, tokenizing string to words, removing punctuation and diacritic marks, applying stop words removal, applying the proper term stemming and pruning methods as feature reduction techniques, normalizing the tokenized words and finally applying the appropriate term weighting scheme to enhance text document representation as feature vectors. We utilize the open-source machine learning tool Rapid Miner for text pre-processing. Table 2 shows all steps of pre-processing for AL-Hadith.

TABLE II
RESULTS OF PRE-PROCESSING PHASE STEPS FOR AL-HADITH

| Step | Result of the step |
|---|---|
| Removing Sanad | أن رسول الله صلى الله عليه وسلم قال حق المسلم على المسلم ست. قيل ما هن يا رسول الله؟ قال: إذا لقيته فسلم عليه، وإذا دعاك فأجبه، وإذا استنصحك فانصح له، وإذا عطس فحمد الله فشمته، وإذا مرض فعده، وإذا مات فاتبعه. |
| Tokenization | {"أن","رسول","الله","صلى","الله","عليه","وسلم","قال","حق","المسلم","على","المسلم","ست",".","قيل","ما","هن","يا","رسول","الله","?","قال",":","إذا","لقيته","فسلم","عليه","،","وإذا","دعاك","فأجبه","،","وإذا","استنصحك","فانصح","له","،","وإذا", "عطس","فحمد","الله","فشمته","،","وإذا","مرض", "فعده","،","وإذا","مات","فاتبعه","."} |
| Removing Punctuation and Diacritic Marks | {"أن","رسول","الله","صلى","الله","عليه","وسلم","قال","حق","المسلم","على","المسلم","ست","قيل","ما","هن","يا","رسول","الله","قال","إذا","لقيته","فسلم","عليه","وإذا","دعاك","فأجبه","وإذا","استنصحك","فانصح","له","وإذا","عطس","فحمد","الله","فشمته","وإذا","مرض","فعده","وإذا","مات","فاتبعه"} |
| Removing Stop Words | {"رسول","الله","صلى","الله","وسلم","قال","حق","المسلم","المسلم","قيل","يا","رسول","الله","قال","لقيته","فسلم","وإذا","دعاك","فأجبه","وإذا","استنصحك", "فانصح","وإذا","عطس","فحمد","الله","فشمته","وإذا","مرض","فعده","وإذا","مات","فاتبعه"} |
| Light Stemming | "فاج,"دعاك","اذا","فسلم","لقيت","قال","له","رسول","يا","قيل","مسلم","مسلم","حق","قال","سلم","له","صل","له",رسول"{" }"فاتبع","مات","اذا"","فعد,"مرض", "اذا","فشمت","له","فحمد","عطس","اذا","فانصح","استنصحك","اذا",ب" |
| Filter Tokens: | {"رسول","له","صل","له","سلم","قال","حق","مسلم","مسلم","قيل","يا","رسول","له","قال","لقيت","فسلم","اذا","دعاك","فاجب","اذا","استنصحك","فانصح","اذا","عطس","فحمد","له","فشمت", "اذا","مرض","فعد","اذا","مات","فاتبع"} |
| Generate 2-Grams | {"رسول","رسول_له","له","له_صل","صل","صل_له","له","له_سلم","سلم","سلم_قال","قال","قال_حق","حق","حق_مسلم","مسلم","مسلم_مسلم","مسلم","مسلم_قيل","قيل","قيل_يا","يا","يا_رسول","رسول","رسول_له","له","له_قال","قال","قال_لقيت","لقيت","لقيت_فسلم","فسلم","فسلم_اذا","اذا","اذا_دعاك","دعاك","دعاك_فاجب","فاجب","فاجب_اذا","اذا","اذا_استنصحك","استنصحك","استنصحك_فانصح","فانصح","فانصح_اذا","اذا","اذا_عطس","عطس","عطس_فحمد","فحمد","فحمد_له","له","له_فشمت","فشمت","فشمت_اذا","اذا","اذا_مرض","مرض","مرض_فعد","فعد","فعد_اذا","اذا","اذا_مات","مات", "مات_فاتبع","فاتبع"} |

In linguistics, morphology is the identification, analysis, and description of the structure of morphemes and other units of meaning in a language like words, affixes, and parts of speech. For the Arabic language, there are two different morphological analysis techniques; stemming and light stemming. Stemming is the process for reducing inflected (or sometimes derived) words to their stem, base or root form – generally a written word form. Stemming algorithm by Khoja [21] is one of the well-known Arabic stemmers. Light stemming, in contrast, removes common affixes from words without reducing them to their stems and keeps the words' meanings unaffected [1, 12, 24]. A light stemmer [23] is a standard Arabic light stemmer.

The aim of term weighting is to enhance text document representation as feature vectors. Popular term weighting schemes are TF-IDF, TF, TO, and BTO. BTO indicates the absence or presence of a word with Boolean 0 or 1 respectively. TF(t,d) is the number that the term t occurred in document d. TO be the number of occurrences of term t in document d. TF-IDF is a weight often used in retrieval and text mining. This weight is a statistical measure used to assess how important a word is to a document in a collection or corpus. Term frequency tf(t, d) is the number that the term t occurred in document d. Document frequency df(t) is the number of documents in which the term t occurred at least once. The inverse document frequency can be calculated from document frequency using the formula: log(num of Docs/num of Docs with word i). A reasonable measure of term importance may then be obtained by using the product of the term frequency and the inverse document frequency (TF*IDF) [12, 20, 24, 27].

## IV. Experimental Results and Evaluation

We perform experiments on Arabic MHAC corpus collected manually from Islam's lawsuit and indicative website [18]. The corpus includes 1,306 text documents belonging to one of the eight categories (the book of Prayers, the book of Zakat, the book of Fasting, the book of Marriage, the book of Transactions, the book of Musaqah, the book Drinks, and the book of Greetings) that chosen from Sahih Muslim. For text classification, we use NB, SVM, and CNB with 10-fold cross-validation. We split the corpus into two parts (90% of the corpus for training and the remaining 10% to test) using stratified sampling, which keeps class distributions remain the same after splitting. We split the corpus in this way to achieve higher classification results.

For assessing the classification results, we use confusion matrices that are the primary source of performance measurement for the classification problem. We have assessed the obtained classification results utilizing the most common classification measures such as accuracy, precision, recall, and F-measure.

The average classification results are depicted in Figure 2. The morphological analysis (stemming, light stemming), term pruning and term weighting schemes (TF-IDF, TF, TO, BTO) have an obvious impact on the classifier performance as shown in Figure 2. The Figure emphasizes that light stemming, and TO representation for CNB classifier has the best classification results (the accuracy is 91.22%, and the F-measure is 91.86%).

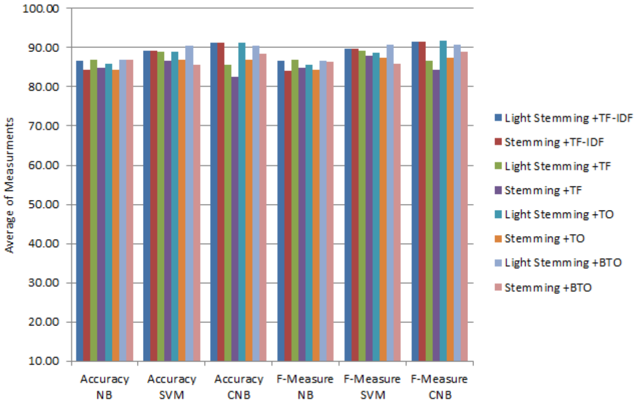

Fig. 2 The classification results for MHAC text representations.

Several observations can be made by analyzing the results in Figure 2. First, using pre-processing techniques like Arabic stop word remover and Arabic stemmer will enhance the accuracy and the F-measure of the classifiers. Second, light stemming has the best classification results this is because lighting stemming is more proper than stemming from linguistics and semantic viewpoint and keeps the word meanings unaffected. Furthermore, classifiers are very sensitive to term weighting schemes because they depend on the distance function to determine the nearest neighbors. For example, the BTO weighting scheme has the worst classification results because the text representation is 0 or 1.

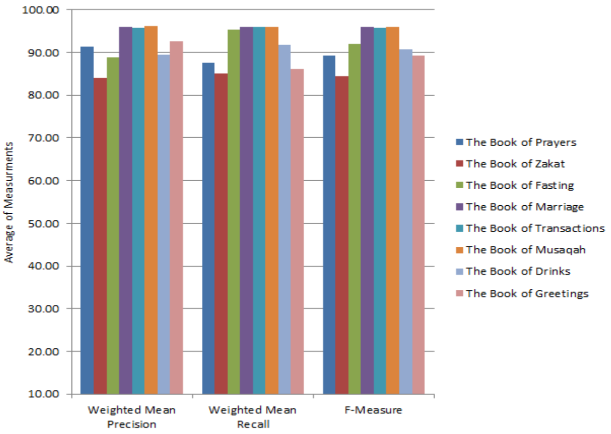

Fig. 3 The classification results for CNB (light stemming + TO)

Figure 3 shows the classification results for the optimal text representation of MHAC corpus (light stemming + TO for CNB) in each of the domain categories. From Figure 2, we can see that the best F-measure is recorded in the book of Musaqah that because the book of Musaqah has limited space of words that are limited and cleared compared with other books. Moreover, it shows that the book of Zakat has the lowest F-measure may be that also because the book of Zakat has a large space domain.

## V. Conclusion and Future Works

This paper studies the impact of text pre-processing and different term weighting schemes on Arabic text classification. In addition, presents and compares the effectiveness of three distinct automatic learning algorithms for classifying Al-Hadith Al-Shareef into eight selective books depending on Sahih Muslim. To the best of our knowledge, there is still no published study on this data set. The classifiers have been tested using Arabic text corpus collected manually by us from the Sahih Muslim, which cover eight books: the book of Prayers, the book of Zakat, the book of Fasting, the book of Marriage, the book of Transactions, the book of Musaqah, the book of Drinks, and the book of Greetings. The learning algorithms are NB, SVM and CNB with 10-fold cross-validation are applied to classify Sahih Muslim Arabic corpus. Moreover, we used TF-IDF, TF, TO, BTO and techniques to compute the relative frequency for each word in a particular document. The results indicate that term stemming and pruning, document normalization, and term weighting dramatically reduce dimensionality, enhance text representation and directly impact text mining performance. Furthermore, classification results show that the CNB achieved promising results compared with other supervised methods in classifying A-Hadith. CNB obtains 91.22% accuracy and 91.86% F-measure.

Possible directions for future work include conducting additional experiments using further text collections to make sure the results that we got. Additionally, we tend to decide to use the other feature choice and weighting methods and compare them with the methods already used. Additionally, enhancing the accuracy of the system, more than one classification method can be merged with each other to increase the accuracy. Finally, it's possible to build a system which can accept as input an archive of texts like Islamic books archive and some category (subject), and as a result, it will give all the texts, which are related to this category.

## References

[1] ABABNEH M., ALNOBANI A., AlSHALABI R., and KANAAN G., "Building an Effective Rule-Based Light Stemmer for Arabic Language to Improve Search Effectiveness", *in Proceedings of the International Arab Journal of Information Technology*, vol. 9, no. 4, pp. 368-372, 2012.

[2] AL-HARBI S., ALMUHAREB A., AL-THUBAITY A., KHORSHEED M. and AL-RAJEH A., "Automatic Arabic Text Classification", *in Proceedings of the 9th International Conference on the Statistical Analysis of Textual Data*, Lyon-, France, 2008.

[3] AL-KABI M. N. and AL-SINJILAWI S., "A Comparative Study of the Efficiency of Different Measures to Classify Arabic Text", *University of Sharjah Journal of Pure and Applied Sciences*, 2007.

[4] AL-KHATIB M., "Classification of Al-Hadith Al-Shareef Using Data Mining Algorithm", *Proceedings of European, Mediterranean & Middle Eastern Conference on Information Systems,* Abu Dhabi, UAE, 2010.

[5] AL-SHALABI R., KANAAN G. and GHARAIBEH M., "Arabic Text Categorization Using KNN Algorithm", *in Proceedings of the 4th International Multiconference on Computer Science and Information Technology,* pp. 5-7, 2006.

[6] BERGER H. and MERKL D., A Comparison of Text-Categorization Methods Applied to N-gram Frequency Statistics, *AI 2004: Advances in Artificial Intelligence, Springer*, pp. 998-1003, 2005.

[7] DUMAIS S., PLATT J., HECKERMAN D. and SAHAMI M., "Inductive Learning Algorithms and Representations for Text Categorization", *in Proceedings of the* 7*th International Conference on Information and Knowledge Management, ACM,* pp. 148-155, 1998.

[8] DUWAIRI R., "Arabic Text Categorization", *in Proceedings of the International Arab Journal of Information Technology*, vol. 4, no. 2, pp. 125-132, 2007.

[9] EL-HALEES A., "Arabic Opinion Mining Using Combined Classification Approach", *in Proceedings of the International Arab Conference on Information Technology (ACIT'2011),* Riyadh, Saudi Arabia, 2011.

[10] EL KOURDI M., BENSAID A. and E-RACHIDI T., "Automatic Arabic Document Categorization Based on the Naive Bayes Algorithm", *in Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages,* Association for Computational Linguistics, pp. 51-58, 2004.

[11] ELKATEB S., BLACK W., VOSSEN P., FARWELL D., RODRÍGUEZ H., PEASE A. and ALKHALIFA M., "Arabic WordNet and the Challenges of Arabic", *in Proceedings of Arabic NLP/MT Conference*, London, UK, 2006.

[12] FELDMAN R. and SANGER J., The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data, *Cambridge University Press*, 2007.

[13] HADI W., THABTAH F., ALHAWARI S. and ABABNEH J., "Naive Bayesian and K-nearest Neighbour to Categorize Arabic Text Data", *in Proceedings of the European Simulation and Modelling Conference.* Le Havre, France, pp. 196-200, 2008.

[14] HAMMAD A., An Approach for Detecting Spam in Arabic Opinion Reviews, 2013.

[15] HARRAG F. and EL-QAWASMAH E., "Neural Network for Arabic Text Classification", *in Proceedings of the 2nd International Conference on Applications of Digital Information and Web, IEEE*, pp. 778-783, 2009.

[16] HARRAG F., EL-QAWASMEH E. and PICHAPPAN P., "Improving Arabic Text Categorization Using Decision Trees", *in Proceedings of the 1st International Conference on Networked Digital Technologies, IEEE*, pp. 110-115, 2009.

[17] HATTAB A. and HUSSEIN A., "Arabic Content Classification System Using Statistical Bayes Classifier with Words Detection and Correction", *in Proceedings of World of Computer Science & Information Technology Journal*, vol. 2, pp. 193, 2012.

[18] Islam website Ministry of Islamic Affairs. April 2014. [Online]. Available:
http://hadith.al-islam.com/.

[19] JBARA K., "Knowledge Discovery in Al-Hadith Using Text Classification Algorithm", *in Proceedings of American Science Journal*, vol. 6, no. 11, 2010.

[20] JING L., HUANG H. and SHI H., "Improved Feature Selection Approach TFIDF in Text Mining", *in Proceedings of the 1st International Conference* on *Machine Learning and Cybernetics, IEEE*, Beijing, pp. 944-946, 2002.

[21] KHOJA S. and GARSIDE R., "Stemming Arabic Text", *Computing Department, Lancaster University*, Lancaster, UK, 1999.

[22] KHREISAT L., "Arabic Text Classification Using N-gram Frequency Statistics a Comparative Study", *in Proceedings of the 2006 International Conference on Data Mining (DMIN'06*), Las Vegas, USA, pp. 78-82, 2006.

[23] LARKEY L., BALLESTEROS L. and CONNELL M., "Light Stemming for Arabic Information Retrieval", *Arabic Computational Morphology: Knowledge-based and Empirical Methods, Springer*, pp. 221-243, 2007.

[24] LEWICKI P. and HILL T., Statistics: Methods and Applications, *Statsoft*, 2006.

[25] RYDING K., A Reference Grammar of Modern Standard Arabic, *Cambridge University Press*, 2005.

[26] SAID D., WANAS N., DARWISH N. and HEGAZY N., "A Study of Arabic Text Preprocessing Methods for Text Categorization", *in Proceedings of the 2nd International Conference on Arabic Language Resources and Tools*, Cairo, Egypt, 2009.

[27] SAUBAN M. and PFAHRINGER B., "Text Categorization Using Document Profiling", *in Proceedings of 7th European Conference Principles and Practice of Knowledge Discovery in Databases (PKDD'03)*, pp. 411-422, 2003.

[28] SEBASTIANI F., "Machine Learning in Automated Text Categorization", *in Proceedings of ACM Computing Surveys (CSUR) Journal*, vol. 34, pp. 1-47, 2002.

[29] SEBASTIANI F., "Text Categorization", *in Proceedings of the Text Mining and its Applications to Intelligence, CRM and Knowledge*, UK, pp. 109-129, 2005.

[30] SONG M. and WU Y., Handbook of Research on Text and Web Mining Techologies, Information Science Reference, USA, 2009.

[31] YANG Y. and LIU X., "A Re-examination of Text Categorization Methods", *in Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 42-49, 1999.

[32] YANG Y., SLATTERY S. and GHANI R., "A Study of Approaches to Hypertext Categorization", *in Proceedings of the Journal of Intelligent Information Systems*, pp. 219-241, 2002.