MJoSHT
Malaysian Journal of Science, Health & Technology
[mjosht.usim.edu.my]

USIM

Article

Generating Arabic Stop-Word for Hadith

Yousef Abd-Almohdi Hazzaimeh[1], Norita Md Norwawi[2], Najm Abdul Rahman Khalaf[3]

[1] *Faculty of Science and Technology, Universiti Sains Islam Malaysia (USIM), 71800 Nilai, Negeri Sembilan, Malaysia*
E-mail: ysf_abd82@yahoo.com

[2] *Islamic Science Institute, Universiti Sains Islam Malaysia (USIM), 71800 Nilai, Negeri Sembilan, Malaysia*
E-mail: norita@usim.edu.my

[3] *Faculty of Quranic and Sunnah, Universiti Sains Islam Malaysia (USIM), 71800 Nilai, Negeri Sembilan, Malaysia*
E-mail: najm@usim.edu.my

***Abstract*—Stop-words or (function words) play a great role in performing various functions in sentences, but are still typically inadequate to use for retrieval. They consist of several elements such as common nouns, pronouns and prepositions. With that, there are several Arabic particles available in the online Khoja and Abu EL-Khair stoplist taken from various websites. Arabic stop-words' main problem lies in accepting the prefixes and suffixes' attachment. In the current paper, a new methodology for generating a general stop-word list has been proposed and applied on hadiths. In detail, hadith is defined as words, acts, deeds, traditions, silent approvals and character of Prophet Muhammad S.A.W. (peace be upon him). The current paper aims at examining the effect of removing stop-words from verification of hadiths. The problem is that the previously generated stop-word lists have been on Modern Standard Arabic (MSA), which is the most commonly used language in hadiths. A stop-word list of Hadith and a corpus-based list has been created to be used in the process of hadith verification. The effectiveness and success of Hadith verification when using the newly generated lists along with earlier generated lists of MSA, combining the Hadith lists have been compared with the MSA lists. The Hadith verification has been performed using sequential pattern mining. Lastly, the experiments have demonstrated that the general lists comprising hadith words showed a better performance compared to using the lists of MSA stopwords.**

***Keywords*— Arabic stop-words; Classical Arabic stop-words; Hadith verification.**

I. INTRODUCTION

Of late, the internet web has turned to be a very significant source of data for being a read-write platform by several users. In fact, many language speakers use the web on a daily basis as it is not only used by English speakers. Arabic, for instance, is articulated by over and above four hundred million people and is considered the "fastest-growing language on the internet web with an annual growth rate of 7,247.3 % in the number of Internet users as of 2017, compared to 3,434.0 % for Russian, 2,650.1 % for Indonesian/Malaysian, 2,286.1 % for Chinese and 599.6 % for English" (https://www.internetworldstats.com/stats7.htm).
Historically, the Arabic language can be traced back to more than 1600 years with an archaic Arabic writing system labelled as the consonantal system. In other words, each single consonant is represented by a letter in the Arabic alphabet. Yet, in the late 7th century, the Arabic diacritics, which are considered graphical symbols that distinguish among the various pronunciations of consonants, have been invented by "Abu Al-Aswad Al-Du'ali".

On the other hand, a lot of users and languages' speakers eliminate them from the current written text. In detail, words with similar form of writing may be discerned via their context by several Arab readers [1]. Both of the Arabic language and Islam simultaneously spread in the Middle East, primarily in the course of the 6th and 7th centuries. In the contemporary linguistic world, as somebody speaks of Arabic, they possibly refer to Modern Standard Arabic (MSA), not Classical Arabic (CA). More specifically, the central difference between MSA and CA lies in the fact that a huge amount of these linguistic terms speaks of certain concepts and items were not existent at the time of CA. Despite the fact that MSA is more likely to be applied to non-Arabic terminologies, CA is mentioned as Quranic Arabic and is considered as the written language of the Holy Quran, the key source of the text of Islam [2].

The considerable amount of data stored in disorganized texts is no longer simple to be adopted for more processing by computers. Consequently, certain algorithms and pre-processing methods are strongly required so as to attain valuable patterns. Stop-words are defined as certain words appearing in the text, carrying little meaning and are commonly used by speakers. Their significance lies only in serving a syntactic function with no indication of any subject matter. As for the IR system, stop-words are defined as "stop-word means high frequency and low discrimination and should be filtered out"[3]. Likewise, stop-word as a concept in the text mining is the same, but its capability to feature a document is closely related to much more significance to

examine if a certain word is a stop-word or not. Yet, it is viewed as not the most applicable practice to enlarge the stop-word list as much as possible, but on the contrary, to raise the recall rate in IR. Concerning the text mining process, the entire words labelled as stop-words shall be precisely elucidated so as to improve the accuracy and effectiveness, thus the fundamental concept has increased the accuracy and effectiveness of the task of the text mining after the stop-words have been filtered out. With this, the following aspects are involved [3]:

1. The text mining's accuracy shall not be reduced in case the non stop-words are deleted.
2. The text feature space dimensionality shall be lessened in case the stop-words are deleted.

Of the most essential steps to a data retrieval system is to recognize a stop-word list that includes words so as to delete them from the text processing. Stop-lists are classified into two categories, namely: "domain independent stop-lists and domain dependent stop-lists". They are formed through the use of three methods, namely: syntactic classes, corpus statistics that is considered a very dependent domain approach used with specific fields or a mixture of both of syntactic classes and corpus statistics to attain the benefits, alongside advantages of the three approaches. More importantly, "whatever the language, there is no definite list of Arabic stop words (stop-list) which all NLP tools incorporate" [3]. More specifically, Arabic is considered a very deep language, which is full of rich lexicons and includes several suffixes and prefixes that may be added to any word to change its meaning into something different. In Arabic language, several letters are applicable for use as prefixes as the word's meaning may be changed. First, the following letters (" أ" "ب" "ف" "ك" ل") are only used in certain words, not all of them.

Secondly, a huge amount of the original words out of the said categories are possibly connected together and used as prefixes and suffixes, particularly pronouns.

Finally, another example to explain the said is the conjunction letters such as WAW meaning "and" and Fa meaning "will" that may be similarly used, but it does not happen because it is possibly used with the entire Arabic words with no exception and it is unrealistic to use it.

The following shall be a brief account of a few of the previously improved stop-lists:

- A quite short stop-list of 168 words used for an Arabic stemmer has been developed by Khoja [4].
- A top-list minted through translating an English list and enhancing it with high frequency words from the corpus leading to a larger 1.131-word list has been developed by Chen and Gey [5].
- Three stop-lists have been created by El-Khair [6]. The first top-list is a general one constructed on the structure of the Arabic language. In detail, the entire articles and papers that are possibly thought to be a stop-word are collected from various syntactic classes in the Arabic language (e.g. Adverbs, Prepositions, Conditional Pronouns, etc) in a more organised way to make sure that the list is truly complete. As a result, there has been a new list with 1,377 words. As for the second list, it is truly a corpus-based list consisting of 359 words noting that these words have occurred more than 2500 in the corpus statistics. The nature of the stop-list is different as it merges the corpus and the general based top-lists together.

Based on the said, the problem lies in the fact that the previously created stop-words are existent on Modern Standard Arabic (MSA), which is the not the most commonly used language in hadiths. Another problem regarding the Arabic stop-words is that they are flexible to accept the prefixes and suffixes of the attachment [2], [6]. The current paper's structure is as follows: Section 2 explains the stop-words' characteristics and description and discusses the elimination techniques of the stop-words. Section 3 examines the research work and studies related to the generation of stop-words, particularly on hadiths. In Section 4, the required new technique to create a list of Arabic stop-words is represented. As for the results, they will be elucidated in Section 5. Finally, the conclusion will be shown in Section 6. This paper presents how a list of general stop-words is generated and compared to other lists, namely: Khoja nd EL-Khair.

II. Related work

In an IR experiment regarding the Arabic language, it is well-known that there is no general standard stop-list to use. As for the stop-list adopted in the Lemur Toolkit, it has been created by Khoja [4]. As she creates her Arabic stemmer with 168 words, this has been used by Larkey [7], [8]. A top-list minted through translating an English list and enhancing it with high frequency words from the corpus leading to a larger 1.131-word list has been developed by Chen and Gey [5]. Yet, they have not addressed the effects of the list.

Moreover, a dependent domain list, which includes three problems, has been created by Savoy and Rasolofo [9]. Firstly, they have used a few words preceded by the letter waw "و" meaning "and" in 17 words together with 11 duplicates. In several words in the Arabic language, this letter comes in a different format and can come before the entire words in the language with no exceptions. One of the most appropriate methods to do this is to eliminate it through the use of an applicable and effective stemming algorithm. Secondly, they have deleted enormous single letters with the waw, specifically "ba' "ب", heh "ه". "hamza "ْ", alef "ا ,أ". Because of the nature of writing the Arabic language, the said letters can come individually, but they are still considered a part of the word as deleting them will change the meaning of the word or make it meaningless, e.g. the word of "كتاب" that has several meanings such as writers, book, or a place of learning includes the letter ba' as single separate letter as removing it will make the word meaningless. Thirdly, few words used in it are not considered stop-words although they have repeatedly appeared in the corpus' statistics' analysis such as Cairo "القاهرة", United "المتحدة", States "الو لايات", etc. Additionally, it is considered a more dependent domain list, thus it is unlikely appropriate for other collections.

Kabi [10] has categorized a group of Arabic hadiths into the so-called "Sahih AL-Bukhari", which is an 8-chapter book. It has been done through having the term frequency calculated. In pre-processing part, the removal of stop-word is done through using an algorithm based on the so–called "deterministic finite machine" [11].

In the meantime, other research works and literary studies have created lists of stop-words, still there has been no general and accepted stop-list or stop-word list for the texts of hadiths. For instance, AL-Kabi [11] recommended doing research work in the future on the impact of steps of pre-processing such as stemming and stop-word removal concerning the results of hadiths. Moreover, other researchers such as Alkhatib [12], Harrag [13], [14], [15], [16], Al-Kabi [17], Saeed and Jaffry [18], Hasan and Zakaria [19] have conducted researches on the Arabic texts of hadiths; however, they have not addressed to what extent a stop-list's use affects their research work.

Also, Harrag and Al-Qawasmah [15] recommended deleting stop-words with high and low frequency words. As for Jbara [20], he has manually helped in building a list of stop-words that includes Arabic prepositions, pronouns, names of people such as Prophet Mohammad's companions as well as places said in the corpus of hadiths. Likewise, Harrag [21] has conducted a different operation that includes the elimination of stop-words required to delete the meaningless definite articles and pronouns from the hadith Matn's text.

Besides, Shatnawi [22] asserted that "it is significant to adopt stop-word and stop-list as the accuracy when extracting hadiths from web pages is of outmost importance. The existence of some common phrases in the text of the Hadith limits the precision of the retrieval. For example the phrase "عن ابي هريرة رضي الله عنه عن النبي صلى الله عليه و سلم", it is a common phrase in more than 1300 hadiths".

It is quite inappropriate to extract all of these 1300 hadiths because one of them is mentioned in the web page. From the experimental point of view, they found any terms that occur more than 750 times in all 17,000 Hadiths should not be indexed and therefore added to the Stop-word list, where the final list included 122 terms.

From the all works reviewed and discussed above we find these questions:

- How to create a general stop-list to be suitable not only for hadith but also to other collection with consideration the different between the CA and MSA?
- How to deal with the attachment of prefixes and suffixes with the consideration that may affect to the meaning or giving ambiguous words?
- Does the dimensionality of the text feature space is reduced after removing these words?
- Does our created list reduce the size better than the list created before such as Khoja and El-Khair stop lists?

In the rest of this paper, these questions will be answered.

III. METHODOLOGY

A. Generate List of All Arabic Stop-Words

A general stop-list was created, based on the Arabic language structure and characteristics without any additions. All possible words or articles that may be considered a stop-word were collected from the different syntactic classes in Arabic in a systematic way to ensure the completeness of the list. Choosing a word based on:

1. They give no meaning if they are used alone.
2. They are general words and not used specifically in a certain field.

The resulting list consisted 128 words, which were collected from Arabic dictionary. These words include:

- attached and separate prepositions
- conjunctions
- interrogative words
- exclamations
- calling letters
- adverbs of time and place
- pronouns
- demonstratives

B. Adding Possible Prefixes to the Words

Arabic is a very rich lexical language, which has a large number of prefixes and suffixes that could be added to a word to change its meaning. The problem facing when dealing with prefix and suffix are, first, the prefix letters like prepositions (" أ" "ب" "ك" "ل ") and suffix letters like pronouns ("ه" "ي " "ا " "ك ") some words when add giving meaning but others not. Example for a word that gives meaning is the word "من", which can accept prefix like "ومن" and suffix like "منه". Another example for word that does not give meaning is the word "كي" which can accept prefix like "وكي" but when any suffix is added on it, it does not give meaning.

Second, there are some of pronouns which can be used alone and can be attached to other word as suffix. For example, word"من" can be attached with the word "هم" like "منهم". Finally, the conjunction letters "و" and "ف" could be used as a prefix for all Arabic words with no exception.

To deal with that, after the words have been collected as mentioned above, the words are grouped into two categories, the first one are the words that accept suffix and the second are the words that do not, after that, all possible suffixes will be added to the first category, but to do this in a systematic way the only words under class as " حروف الجر ما يشترك بين الظاهر والمضمر" which translate to English as "Prepositions which is shared between the apparent and the incarnate" be in this category, which contained 11 words. Then all words will have to be combined, finally the conjunction letters WAW and Fa will be added to all words to complete the list. These letters can be added to any words as prefix without any exception. Fig. 1 shows the steps to create the final list.

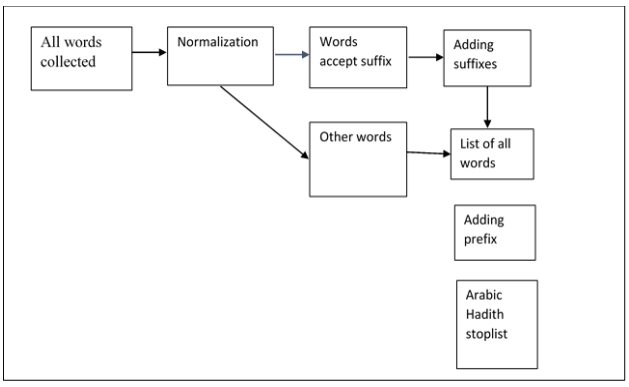

Figure 1: Steps for creating the general stop list

IV. Results and Discussion

The experiment used a corpus which contains 14654 hadiths, which from Jaami Al-Sagheer and Daeef Al-Jami Al-Sagheer. Hadith texts classified as sahih (sound), hasan (good), da'if (weak), da'if jiddan (very weak) and maudu' (fabricated) are tokenized to words, and the non-Arabic alphabets were removed after the normalization step. In normalization the following techniques were used:

- The letters (آ ، أ، إ) were replaced with (ا).
- The letter (ؤ) was replaced with (و).
- The final (ى) was replaced with (ي).
- The Final (ة) was replaced with (ه).

The books of Sahih and Daeef Al-Jami Al-Sagheer give 231574 words. After combining them together and removing the duplicates, the list of all words is 201031with 33604 distinct words. Then we have calculated the frequency of occurrence of each word from the list of all words in the Sahih and Daeef Al-Jami Al-Sagheer combined together. Table 1 and 2 show the top ten word's frequency before and removing the duplicates.

TABLE 1
Top Ten Word's Frequency Before Removing The Duplicates

Words	Count
من	8687
الله	5687
في	4981
ان	4069
لا	3144
على	2597
ما	2535
الا	2276
اذا	2063
ولا	1922

TABLE 2
Top Ten Word's Frequency After Removing The Duplicates.

Words	Count
من	5684
الله	3971
في	3455
ان	3244
لا	2590
على	2052
ما	1991
اذا	1927
الا	1802
كان	1521

Table 1 shows that the nine of ten words are function words. Also, the most seven frequent words are the same in Table 2. The 8th and 9th most frequent word are changed after the duplicates have been removed, where the 10th most frequent word is totally changed.

The list we collected contain 128 words, 11 of them are related to the first category, where all of those are the preposition words. Those words will generate all suffixes first, and then the collection contains 129 words. After that the two categories will be combined, which contain 246 words to add with the conjunction letters on it, the final list contains 738 words. Table 3 shows the top ten words after removing the stop-words.

TABLE 3
Top Ten Word's Frequency After Removing The Stopwords.

Words	Count
الله	3971
يوم	1059
الجنه	917
تعالى	820
الناس	776
القيامه	664
احدكم	644
النار	567
اهل	468
الرجل	451

Table 3 shows that the ten words after removing the stop-word are not function words, that was the reason why we delete only function words because the hadith is very important text and the accuracy should not be decreased if the stop-words were deleted.

The words chosen as a stop-word to be removed from the corpus are general words, which means these words did not have any effect on the accuracy of data. To evaluate the effectiveness of the new Arabic stoplist, we compare the works of the general stoplist with Elkair general stoplist and khoja stoplist Fig. 1, 2 and 3 show the main difference between them.

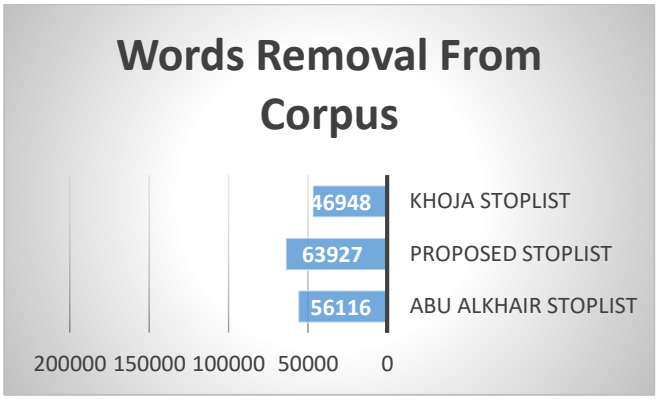

Figure 1: Number of Words Removed from Corpus

Fig. 1 shows that the proposed stop-list removed 63927 from the corpus with 56116 for Abu Alkhair and Khoja stop-list at last with 46948.

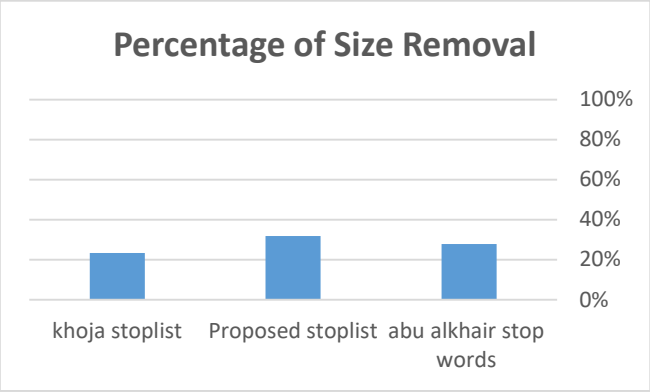

Figure 2: Percentage of Size Removal

Fig. 2 shows that the proposed stoplist removed 32% of the corpus with 28% for Abu Alkhair and Khoja Stoplist at last with 23%.

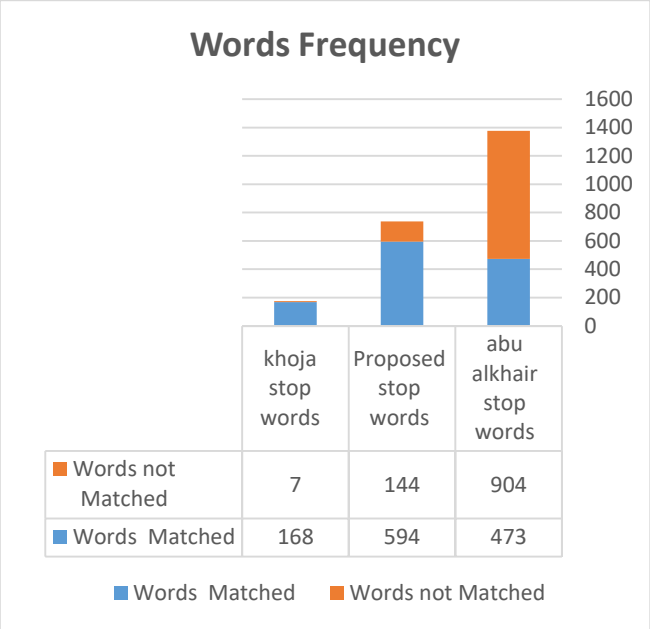

Figure 3: Words Frequency

Fig. 3 shows that the words in the proposed stop-list were matched from the list by 594 compared with 473 for Abu Alkhair and Khoja's stop-list at last with 161.

V. Conclusions

In the text mining process, stop-word should be filtered out to improve the efficiency and accuracy. Stop-words are common words that generally do not contribute to the meaning of a sentence. The common English words that do not affect the meaning of a sentence are like "a", "the", "of" and etc. The case is similar in Arabic, where the common Arabic words that do not affect the meaning of a sentence are like "من", "في", "الى", so removing the stop-words will reduce the corpus size without losing important information. One of the problems of Arabic stop-words is that they accept the attachment of prefixes and suffixes which is not in English. Arabic is a very rich lexical language which has a large number of prefixes and suffixes that could be added to a word, therefore, these letters can change the meaning of the word when attach on it. In another word, due to the way the Arabic language is written, these letters can come separately, but they are still a part of the word and removing those will change the word meaning or leave it meaningless. To solve this problem, we collect all common words in Arabic first, then divided them into two categories. In the first category the prepositions which is shared between the apparent and the incarnate contain 11 words. In the second category, the rest of the words contain a total of 117 words. After adding all suffixes to the first category the number of words decreased to 129 words. After adding the conjunction letters to all the words, the number of words decreased to 738 words. The evaluation on the Corpus shows that the superiority of the performance in terms of the number of stop-words removed.

In this article, we have proposed a generalization of the content of the resource and proposed a standard structure for its representation. In the future, we plan to try more text processing techniques on Arabic hadith data like the verification technique.

References

[1] D. E. M. Abuzeina, "Utilizing data-driven and knowledge-based techniques to enhance Arabic speech recognition," King Fahd University of Petroleum and Minerals (Saudi Arabia), 2011.

[2] W. Medhat, A. H. Yousef, and H. Korashy, "Egyptian dialect stopword list generation from social network data," *arXiv preprint arXiv:1508.02060*, 2015.

[3] Z. Yao and C. Ze-Wen, "Research on the Construction and Filter Method of Stop-word List in Text Preprocessing," *2011 Fourth International Conference on Intelligent Computation Technology and Automation*, 2011.

[4] S. Khoja and R. Garside, "Stemming arabic text," *Lancaster, UK, Computing Department, Lancaster University*, 1999.

[5] A. Chen and F. Gey, "Translation term weighting and combining translation resources in cross-language retrieval," in *TREC*, 2001, p. 2001: Citeseer.

[6] I. A. El-Khair, "Effects of stop words elimination for Arabic information retrieval: a comparative study," *arXiv preprint arXiv:1702.01925*, 2017.

[7] L. S. Larkey and M. E. Connell, "Arabic information retrieval at UMass in TREC-10," MASSACHUSETTS UNIV AMHERST CENTER FOR INTELLIGENT INFORMATION RETRIEVAL, 2006.

[8] L. S. Larkey, L. Ballesteros, and M. E. Connell, "Improving stemming for Arabic information retrieval," *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR 02*, 2002.

[9] J. Savoy and Y. Rasolofo, "Report on the TREC 11 experiment: Arabic, named page and topic distillation searches," in *TREC*, 2002.

[10] M. N.Al-Kabi, G.Kanaan, R.Al-Shalabi, S. I. Al- Sinjilawi, and R. S. Al- Mustafa, "Al-Hadith Text Classifier," *Journal of Applied Sciences*, vol. 5, no. 3, pp. 584–587, 2005.

[11] M. N. Al-Kabi, H. A. Wahsheh, and I. M. Alsmadi, "A topical classification of hadith Arabic text," *IMAN*, vol. 2014, pp. 2nd, 2014.

[12] M. Alkhatib, "Classification of Al-Hadith Al-Shareef using data mining algorithm," in *European, mediterranean and middle eastern conference on information systems, EMCIS2010, Abu Dhabi, UAE*, pp. 1-23, 2010.

[13] F. Harrag, A. Hamdi-Cherif, and E. El-Qawasmeh, "Vector space model for Arabic information retrieval — application to 'Hadith' indexing," *2008 First International Conference on the Applications of Digital Information and Web Technologies (ICADIWT)*, 2008.

[14] F. Harrag, A. Hamdi-Cherif, A. M. S. Al-Salman, and E. El-Qawasmeh, "Experiments in improvement of Arabic information retrieval," in *3rd International Conference on Arabic Language Processing (CITALA), Rabat, Morocco*, pp. 71-81, 2009.

[15] F. Harrag and E. Al-Qawasmah, "Improving Arabic Text Categorization Using Neural Network with SVD," *JDIM*, vol. 8, no. 4, pp. 233-239, 2010.

[16] F. Harrag, E. El-Qawasmah, and A. M. S. Al-Salman, "Stemming as a feature reduction technique for arabic text categorization," in *10th*

International Symposium on Programming and Systems, pp. 128-133, 2011.

[17] M. N. Al-Kabi, H. A. Wahsheh, I. M. Alsmadi, and A. Al-Akhras, "Extended topical classification of hadith Arabic text," *Int J Islam Appl Comput Sci Technol*, vol. 3, no. 3, pp. 13-23, 2015.

[18] A. R. Saeed and S. W. Jaffry, "Information Mining from Islamic Scriptures," in *Proceedings of the 4th Workshop on South and Southeast Asian Natural Language Processing*, pp. 66-71, 2013.

[19] A. M. Hasan and L. Q. Zakaria, "QUESTION CLASSIFICATION USING SUPPORT VECTOR MACHINE AND PATTERN MATCHING," *Journal of Theoretical & Applied Information Technology*, vol. 87, no. 2, 2016.

[20] K. Jbara, "Knowledge discovery in Al-Hadith using text classification algorithm," *Journal of American Science*, vol. 6, no. 11, pp. 409-419, 2010.

[21] F. Harrag, "Text mining approach for knowledge extraction in Sahîh Al-Bukhari," *Computers in Human Behavior*, vol. 30, pp. 558–566, 2014.

[22] M. Q.shatnawi, Q. Q. Abuein, and O. Darwish, "Verifying Hadith Correctness in Islamic Web Pages using Information Retrieval Techniques," *International Journal of Computer Applications*, vol. 44, no. 13, pp. 47–50, 2012.