DOI: http://dx.doi.org/10.26483/ijarcs.v9i5.6301

ISSN No. 0976-5697

Volume 9, No. 5, September-October 2018

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

STOPWORDS REMOVAL AND ITS ALGORITHMS BASED ON DIFFERENT METHODS

Jashanjot Kaur
Department of Computer Science and Engineering
Sant Longowal Institute of Engineering and Technology
Longowal, Sangrur (Punjab) – 148106, India

Preetpal Kaur Buttar
Department of Computer Science and Engineering
Sant Longowal Institute of Engineering and Technology
Longowal, Sangrur (Punjab) – 148106, India

Abstract: In this paper analysis of different methods to remove stopwords in Punjabi language have been done. For organizing unstructured text in order to implement stopwords removal techniques, text preprocessing has to be applied. Text processing describes a variety of processing that is performed on raw data to prepare it for one more processing procedure which will be more helpful for performing some further, more purposeful analytic tasks. The words are called stopwords that occur most frequently in a document and contain very little information which is not essential in a document such as ਦੇ, ਹੈ, ਦੀ, ਤੇ, ਦਾ, ਨੇ, ਅਤੇ, ਤੋਂ etc. A list of such words is known as 'stopwords list' or 'stopwords corpus'. These words are removed in the preprocessing phase of the text classification process. The process of removing stopwords help to save time and reduces the size of the document. It also helps to increase performance of information retrieval (IR) tasks. Most of the researchers worked on languages such as English, Arabic, Sanskrit etc. in the IR field. Therefore a lot of work and efforts need to be done in languages other than the languages in which the research has been already done to a great extent. The main goal of this paper is to remove the stopwords in Punjabi language by using different techniques. Punjabi language is 11th most-spoken language of India which is written in Gurmukhi script. Punjabi language is also used in mass media such as news, advertisements, movies, music etc. There is no standard stopword list created for Punjabi language as most of the stopwords lists are created for English and other languages. In this paper, four different methods and its algorithms i.e Classic method using a pre-compiled stoplist, method based on frequency, method based on removing singletons, method based on Punjabi words corpus have been proposed, implemented and the results are analyzed to remove Punjabi stopwords. Thus the size of the document is reduced by 20-30% by eliminating the set of such stopwords.

Keywords: Stopwords removal techniques; stopwords corpus; Classic method based on pre-compiled stoplist; method based on frequency; method based on removing singletons; method based on Punjabi words corpus

I. INTRODUCTION

The main purpose of the research work is to keep the useful data by discarding the unuseful data. Such unuseful data are known as stopwords. Stopwords were first introduced in 1958 by H.P. Luhn [1]. A short report may have countless typographical components, phrases, layout artifacts etc. Therefore it takes a lot of time to look through these documents. In order to save the amount of time, it is needed to eliminate the noise from the text and focus only on the words that are important in a text. Noise is a superfluous content that is not applicable to the task that needs to be done [2]. Stopword is one of the example of noise in any given information (useful words and other basic words of any language that more often than not added to the semantics of the reports and have no perused included esteem [3]. Eliminating stopwords will not only save time but also reduces the size and vector space of the text. In many natural language processing applications, an efficient stop-word extraction technique is required such as in information retrieval systems (IR), stem weighting, stemming and spelling standardization [4]. The main goal of this research is to remove stopwords in Punjabi language texts. The Punjabi words such as □□, □□, □□, □□□, □□, □□□□, □□, □□, □□□, □□ etc are some of the most common words that occur most frequently in Punjabi text. Stopwords list is a set of stopwords which is also known as 'stopwords corpus'. However, there is no standard stopwords list created for Punjabi language as most of the stopwords lists are created for English and other languages.

II. RELATED WORK

Many researchers have undertaken the task of stopword removal by suggesting some methods for finding the stopwords. A lesser amount of work has been done in other languages comparatively to the research done in English language.

C. Fox, 1990 [5] created a stop list based on the Brown corpus. About 1,014,000 words were taken from English literature. As a result the list of 421 stop words was achieved which is more efficient and effective in order to filter most frequently occuring words as well as the words that were semantically neutral in general English literature.

R. B. Myerson, 1996 [6] proposed that a word to be a stopword must fulfill two conditions. First, a word must have high document frequency (DF). Second, there must be small statistical correlations with all the categories of classification. The statistical correlation between a word and categories of classification was measured by $\chi 2$ (weighted Chi-squared statistic). The Chinese corpus of the Mayor's Public Access Line Project texts had been taken for comparing the results of classifiers after removing the stopwords. He concluded that the creation of stopword list containing 500 words decreased the words by 43% in a corpus and the micro-average was improved nearly 7% from 81.39% to 88.76%.

A. Alajmi et al., 2012 [3] proposed information statistics and a statistical approach to extract an Arabic stopwords list. The comparison of this extracted list is done with a general list. This comparison also showed that the top 200

words of generated list outperformed the general list with a 96% efficiency of the classifier, versus 90% when using the generalized list. Two aspects were involved to achieve better efficiency and accuracy after the expulsion of stopwords in the text mining task i.e. the accuracy of text mining should not be decreased and the dimensionality of the text feature space should be reduced if the stopwords were deleted. This Arabic stop-word list was based on word frequency calculation, mean and variance calculation, entropy calculation and aggregation. The authors had used a corpus containing 1000 documents. The collection of these documents contained more than 700,000 words which included 140781 unique words.

Yuang et al., 2012 [7] evaluated the possibility of enhancing the performance of a statistical machine translation system. The terms that were removed were inserted back in the relaxed output by utilizing a n-gram based word predictor. The data used were taken from the EPPS (Europarl Corpus) which involves 1.2M sentences with 35M words in the training set and 2K sentences each in the development set and test sets. Therefore the results showed that predicting the 40% most frequent words gives 77% accuracy in the text. Therefore the perplexity reduces and did not provide better translations.

Asubiaro and Victor, 2013 [8] proposed entropy based algorithm to identify the possibility of stopwords for Yoruba Language. Two arrangements of corpus of 756,039 Yoruba words were used; the diacritized and undiacritized versions. All words whose entropy was more prominent than 0.6 but were not nouns were considered as a stopwords. A stopwords list of 256 words was drawn from the diacritized texts while a stopword list of 189 words was drawn from the undiacritized texts. The removal of the stopwords reduced the full text by 65.91% in the diacritized texts and 67.46% in undiacritized texts. Only 69.1 % of the stopwords had corresponding words in English stoplist. The author's discoveries suggested that current English stoplist will not work ideally and could not be used for Yoruba language.

Sadeghi et al., 2014 [9] proposed a successful approach to achieve light stop word list in Persian language by recognizing them with the high-frequency terms, normalized inverse document frequency and information model. The stopwords represent tokens that reduce the size of the indexing structure. For example, the 20 top Persian stop words reduce the size of index terms by 22%. Similarly 32 Persian stopwords reduce the size of index terms by 27%. This approach had been applied on different corpora to remove stopword list by method called aggregation in order to recognize Persian stopword list.

Vandana Jha et al., 2016 [10] proposed an algorithm based on Deterministic Finite Automata (DFA) for removing stopwords for Hindi Language. About 200 documents were taken in order to test this algorithm. This algorithm took 1.77 seconds to remove the stopwords and achieved 99% accuracy.

III. PROPOSED METHODOLOGY

Creation of Punjabi Stoplist: A list of Punjabi stopwords was prepared for general texts. The list was prepared by taking the most common words which are of little value that help to select documents according to the need of the user are therefore, entirely removed from the document. The list contains about 220 words that are drawn from wide range of Punjabi literature. The list prepared is as follows:

ਹਨ	□□	□□□□	ਵੀ
□□□	□□□□	□□□	ਕੋਈ
□□□□	□□□	□□□	'ਤੇ
□□	□□	□□	ਤਾਂ
□□□	□□□□	□□□	ਹੋਣੀ
□□□□	□□□	□□□	ਹੁਣੇ
□□□	□□□□□□	□□□□□□	ਚਾਹੀਦੇ
□□□□□□□	□□□□□□□	□□□□□□□	ਤਰ੍ਹਾਂ
□□□□	□□□□	□□□□□□	□□□
□□	□□	□□	□□□□
□□□□	□□□□	□□□□	□□
□□	□□	□□□	□□□
'□	□□	□□	□□□
□□	□□	□□□	□□□
□□	□□	□□	□□
□□	□□	□□□□	□□□
□□		□□□	□□□□
□	□□	□□	□□□□□□
□□□□□□	□□□□	□□□□	□□
□□	□□□	□□	□□□□
□□	□□□□	□□□□	□□□□□□
□□□□	□□□□	□□	□□
□□□	□□□	□□□□	□□
□□□	□□□	□□□□□	□□□□
□□□	□□□□□	□□□□□	□□□□□
□□□□□□□	□□□□	□□□□	□□□□
□□□□□□	□□□□□	□□□□□□	□□□
□□	□□	□□	□□□
□□□	□	□□	□□□
□	□□	□	□□□□
□□□□	□□□□	□□□□□	□□□□
□□	□□□□	□□□□□	□□□
□□□□	□□□□□	□□□	□□□
□□□□□	□□□□□	□□□□□□□	□□□□□
□□□□□	□□□□	□□□	□□□
□□□□	□□□	□□□□□	□□
□□□□□□	□□□□□□	□□□□	□□□
□□	□□□□	□□□	□□□□
□□□□□□	□□□□	□□□□□	□□□
□□□	□□□□□	□□□□	□□□□□□
□□□□	□□□□	□□□□□□	□□□□
□□□□	□□□□□□	□□□□	□□□□
□□□	□□□□	□□□□□	□□□□
□□□□□□	□□□□	□□□□	□□□□
□□□	□□□□	□□□□□□	□□□□□□
□□□□	□□□□□	□□□□□	□□□□□
□□□□□□□	□□□□	□□□□□□	□□□□
□□□□	□□□□□□	□□□	□□□
□□□□□	□□□□□	□□□□□	□□□□□

□□□□□□□	□□□□□	□□□□□	□□□□
□□□□	□□□□□□	□□□□□	□□□□□
□□□□□□□	□□□□	□□□□	□□□□□□
□□□□□□	□□□□	□□□□□	□□□□
□□□	□□□□	□□□□	□□□□

Corpus for the study

A representative and qualitative corpus is taken for this research. Punjabi words corpus is a list of around 65,000 Punjabi words prepared from different text articles written in Punjabi language from different sources such as newspapers, books, magazines etc. The corpus is organized as a list of words along with their respective frequency of occurrence. The list is sorted in decreasing order of frequency with the most frequent words at the top. The stoplist is also compared with the Punjabi corpus as shown in the table below:

Table I. Comparison of stoplist with the Punjabi corpus

S.No.	Top k most frequent words selected from Punjabi Corpus	No. of words in the stoplist matching top k most frequent words in Punjabi Corpus	No. of words remaining in Punjabi corpus	%age of stopwords removed
1.	k= 10	9	1	90%
2.	k= 20	18	2	90%
3.	k= 30	23	7	77%
4.	k= 50	35	15	70%
5.	k= 70	45	25	64%
6.	k= 80	49	31	61%
7.	k= 100	59	41	59%
8.	k= 125	71	54	57%
9.	k= 150	78	72	52%
10.	k= 200	111	89	56%

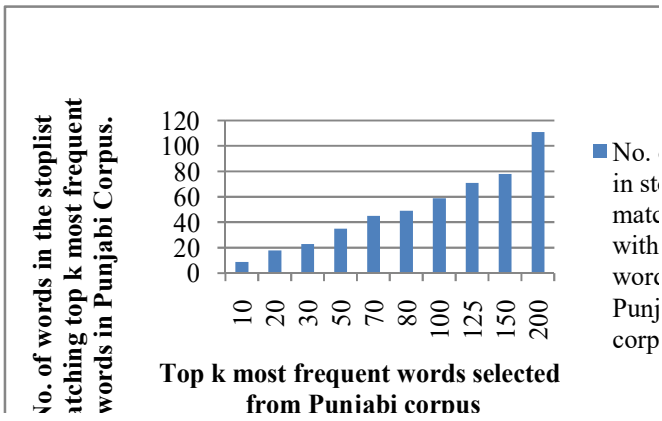


Figure 1. Comparison of stoplist with the punjabi corpus

In this section the different methods and its algorithms for the removal of stopwords in Punjabi language have been proposed, implemented and analyzed. A Punjabi text of approximately 1000 words has been taken. As a preprocessing step, all the special characters, digits and punctuation marks are removed from the input document. The methods are listed below:

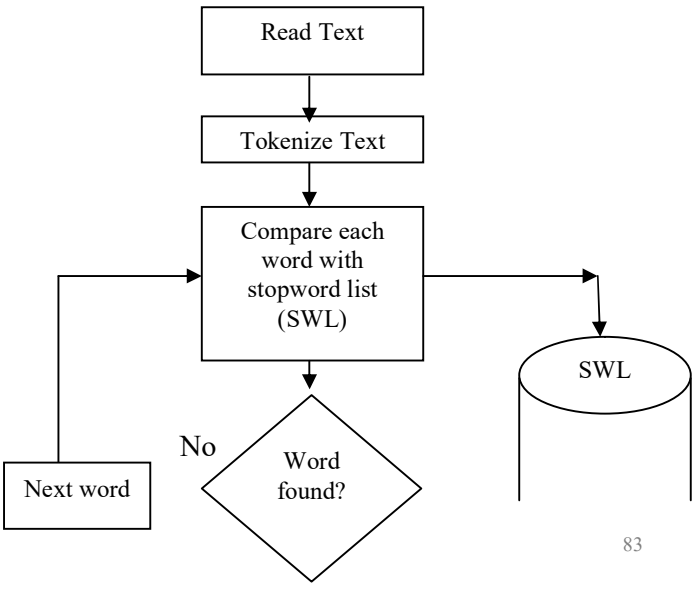
A. The Classic Method

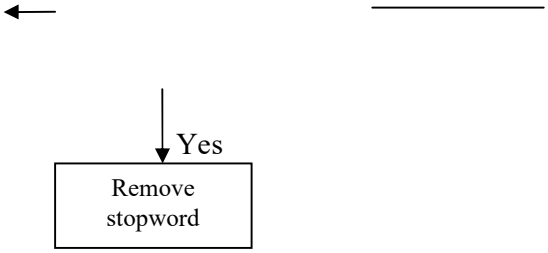
In Classic method, the stopwords are removed by comparing the given text with the Punjabi stoplist.

An algorithm for the classic method is defined below:

Step 0: Input: A text containing words in Punjabi language.
Output: Text after the removal of stopwords.
Step 1: Read the words from the text.
Step 2: Tokenize the words in the given text.
Step 3: For each word w_i :
Compare w_i to all the words in the stoplist
If match occurs:
Remove w_i from text
Step 4: Read the next word.

Flow Chart:-





The following table shows the reduction in size of input text after applying the classic method.

Table II. Analysis of Classic method

S.No.	Document	Categories	No. of stopwords removed	Size reduced
1.	**Books**			
(1)		Book 1	381	39%
(2)		Book 2	397	39%
(3)		Book 3	428	42%
(4)		Book 4	490	42%
2.	**Stories**			
(1)		Story 1	472	47%
(2)		Story 2	353	35%
(3)		Story 3	400	40%
(4)		Story 4	317	32%
3.	**News**			
(1)		Ajit	294	29%
(2)		Jagbani	364	36%
(3)		Punjabi Tribune	335	33%
(4)		Rozana Spokesman	332	32%
4.	**Articles**			
(1)		Ajit	383	39%
(2)		Jagbani	397	40%
(3)		Punjabi Tribune	442	43%
(4)		Rozana Spokesman	345	40%
5.	**Poems**			
(1)		Poem1	301	30%
(2)		Poem 2	352	35%
(3)		Poem 3	356	37%
(4)		Poem 4	438	44%

Observations: The most frequently occurring words in the text i.e. stopwords are removed from the text which reduces about 38% size of the text on an average.

B. Method based on frequency

In this method, the frequency of each distinct word in text is calculated and then the words are sorted in decreasing order of their frequencies. Then top k% most frequent words are removed from the text as stopwords.

An algorithm for the method based on frequency is defined below:

Step 0: Input: A text containing words in Punjabi language.
Output: Text after the removal of stopwords.
Step 1: Read the words from the text.
Step 2: Tokenize the words in the given text.
Step 3: Calculate the frequency of each distinct word.
Step 4: Sort the words according to frequency f_i in decreasing order
Step 5: Remove the top k% most frequent words (k = 1, 2, 3……..)

Flowchart:-

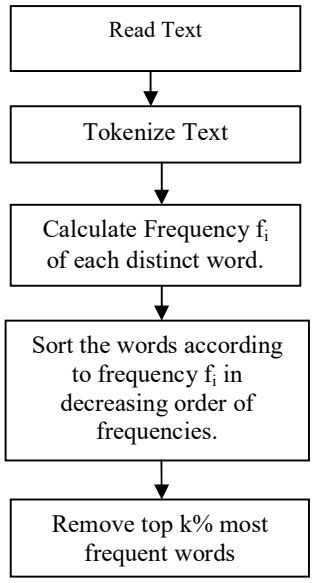


Table III. Analysis of Frequency-based method

Top k most frequent distinct words selected from text.	Total no. of words in top k% most frequent distinct words	No. of Distinct words	No. of stopwords removed	%. of stopwords removed	Size reduced
k = 1%	128	5	5	100%	13%
k = 2%	192	10	10	100%	19%
k = 3%	242	16	13	81%	24%
k = 4%	280	21	17	81%	28%
k = 5%	314	26	20	76%	31%
k = 6%	343	31	22	72%	34%
k = 7%	369	36	24	65%	40%
k = 8%	395	42	26	62%	40%
k = 10%	437	52	29	56%	44%

Observations:-

1. Top 1% to 4% most frequent words in a text contain about 81% stopwords.
2. Removing top 1% to 4% most frequent words reduces the size of the text by 13% to 28%.

C. Method based on removing singletons

In addition to the frequency-based stopword removal method, removal of words that occur once i.e. singleton words has been done. In this method, the frequency of stopwords is calculated and then the words with frequency 1 are removed. For example, in our experiments, the singletons include words such as ਗੀਤ, ਖਾਦਾ, ਦੂਜੇ, ਗਾਉਣ etc. which might play an important role in IR tasks.

An algorithm for the method based on removing singletons is defined below:

Step 0: Input: A text containing words in Punjabi language.

Output: Text after the removal of stopwords.

Step 1: Read the words from the text.

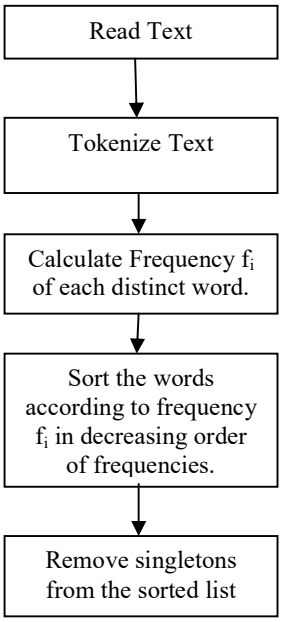
Step 2: Tokenize the words in the given text.

Step 3: Calculate the frequency of each distinct word.

Step 4: Sort the words according to frequency f_i in decreasing order

Step 5: Remove singletons from the sorted list.

Flowchart:



When the singletons present in each text were compared with our stoplist, we found that the percentage of stopwords in the singletons is around 3% on average which means the singletons are mostly information carrying words and their removal can adversely affects the IR task as shown in the following table:

Table IV. Method based on removing singletons

Text	Categories	Total no. of words in the text.	Total No. of Singletons in a text	No. of singletons matching the words in the text.	% of singletons removed	Size Reduced
Books						
	Book 1	986	430	28	7%	3%
	Book 2	1012	348	28	8%	3%

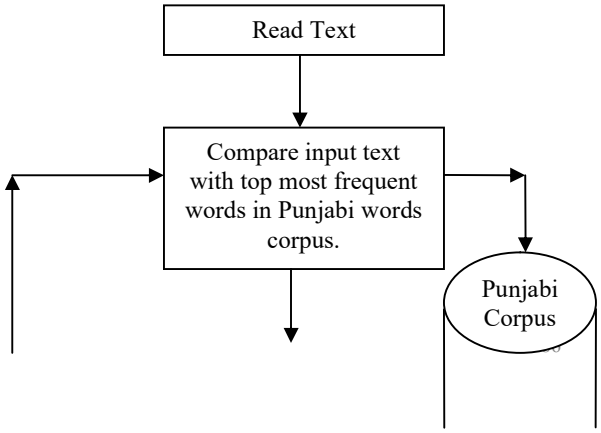
	Book 3	1010	328	32	10%	3%
	Book 4	1012	288	23	8%	2%
Stories						
	Story 1	1007	246	17	7%	2%
	Story 2	1010	385	36	9%	4%
	Story 3	1000	348	27	8%	3%
	Story 4	1000	400	29	7%	3%
News						
	Ajit	1005	338	22	7%	2%
	Jagbani	1000	338	21	6%	2%
	Punjabi Tribune	1016	263	17	6%	2%
	Rozana Spokesman	1031	318	24	8%	2%
Articles						
	Ajit	978	349	40	11%	4%
	Jagbani	1001	387	27	7%	3%
	Punjabi Tribune	1031	332	20	6%	2%
	Rozana Spokesman	863	288	24	8%	3%
Poems						
	Poem1	1016	374	21	6%	2%
	Poem 2	1009	465	23	5%	2%
	Poem 3	966	366	22	6%	2%
	Poem 4	1002	353	25	7%	2%

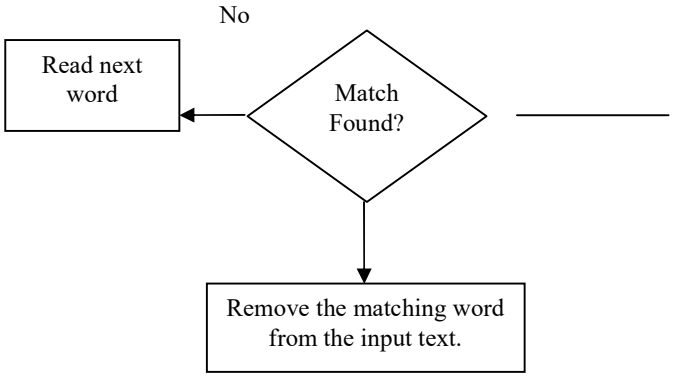
Observations:-
The least frequently occurring words in the text i.e. the singletons are not the best candidates to be removed as stopwords. Infact, the singletons carry significant information content.

D. Method based on Punjabi words corpus

In this method stopwords are removed by comparing the text with Punjabi words corpus. The input text is compared with the top k most frequent words in the Punjabi words corpus word by word. The matching words are then treated as stopwords and thus removed from the input text.
The efficient algorithm for the proposed work is:
Step 0: Input: A text containing words in Punjabi language.
Output: Text after the removal of stopwords.
Step 1: Read the words from the text.
Step 2: For each word w_i
Compare input text with top k most frequent words in Punjabi words corpus.
If match occurs:
Remove matching word from the corpus.
Step3: Read the next word.

Flowchart:





Method based on Punjabi words corpus

S.No.	Top k most frequent words selected from Punjabi corpus	No. of words in the text matching top most frequent words in Punjabi Corpus and thus removed	Size reduced
1.	k = 10	95	8%
2.	k = 20	150	13%
3.	k = 30	171	15%
4.	k = 50	210	13%
5.	k = 70	243	22%
6.	k = 80	248	22%
7.	k = 125	289	26%
8.	k = 150	303	27%
9.	k = 200	322	29%

Observations:
Top 10 to 50 most frequent words selected from Punjabi corpus matches with 95 to 210 words in a text that are considered as stopwords and are removed which reduces the size of the text by 8% to 13%.

IV. CONCLUSION

Stopwords are the no-information words which do not contribute any information to the text-processing task. Instead, they may degrade the performance if included in the text processing because there are only a small number of stopwords that constitute a large fraction to the total text. It is therefore required to eliminate the stopwords during the preprocessing phase. In this work, algorithms to improve Punjabi stopwords are proposed and implemented using various methods i.e. classical method, method based on frequency and removing singletons and a method based on Punjabi corpus. A stoplist of about 220 words is prepared. For experimental setup, 4 News articles, 4 Poems, 4 Stories, articles from 4 different Books, 4 Articles were taken from different Punjabi newspapers such as Ajit, Punjabi Tribune, Rozana Spokesman, Jagbani etc with average of approximately 1000 words each. These methods improve overall performance and execution time. These methods also reduce the size of the text by 20-30%.

V. FUTURE SCOPE

A number of stopword removal methods have been developed by the researchers in the past, particularly for the English language. There is a requirement of efficient stopword removal techniques to be developed for other languages also. Experiments with removing stopwords in Punjabi language is still a new area of research. Therefore it still requires more searching and exploring. The algorithms proposed in this work can also be applied to other languages in future. Further investigation can be done to remove stopwords by proposing new methods and algorithms. Removing stopwords is one of the pre-processing phase. There are more phases to explore in preprocessing such as normalization of nouns in noun morph, allowing input restrictions to input text, lexical analysis etc. List prepared in this work can also be modified to enhance results.

VI. REFERENCES

[1] H. P. Luhn, "A Statistical Approach to Mechanized Encoding and Searching of Literary Information," IBM J. Res. Dev., vol. 1, no. 4, pp. 309–317, 1957.

[2] R. Nisbet, J. elder, G. Miner, "Handbook of statistical analysis and data mining applications", academic Press, Elsevier, 2009.

[3] A. Alajmi, E. M. Saad, R. R. Darwish, "Toward an ARABIC Stop-Words List Generation", International Journal of Computer Applications (0975 – 8887) Volume 46– No.8, May 2012.

[4] B. Alhadidi and M. Alwedyan, "Hybrid Stop-Word Removal Technique for Arabic Language.," Egypt Comput Sci, vol. 30(1), no. 1, pp. 35–38, 2008.

[5] C. Fox, A stop list for general text. ACM- SIGIR Forum, 24, 19-35, 1990.

[6] R. B. Myerson, "Fundamentals of social choice theory", Discussion Paper No.1162, 1996.

[7] C. T. Yuang, R. E. Banchs, C. E. Siong, "An Empirical Evaluation of Stop Word Removal in Statistical Machine Translation", Journal of Computational Linguistics, 30-37, 2012.

[8] Asubiaro, T. Victor, " Entropy-Based Generic Stopwords List for Yoruba Texts", International Journal of Computer and Information Technology (ISSN: 2279 – 0764) Volume 02– Issue 05, September 2013.

[9] M. Sadeghi, J. Vegas, "Automatic identification of light stopwords for Persian information retrieval systems", Journal of Information Science, 1–12, 2014.

[10] V. Jha, Manjunath N, P. Deepa Shenoy and Venugopal K R, "Hindi Stopword Removal Algorithm" International Conference on Microelectronics, Computing and Communication, MicroCom, 2016.

DOI: http://dx.doi.org/10.26483/ijarcs.v9i5.6301

ISSN No. 0976-5697

Volume 9, No. 5, September-October 2018

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info