Building Hybrid Stop-Words Technique with Normalization for Pre-Processing Arabic Text

Jaffar Atwan[†*]

*Corresponding author E-mail: jaffaratwan@bau.edu.jo (Jaffar Atwan)

[†]CIS, Al-Balqa Applied University, Jordan

Summary

In natural language processing, commonly used words such as prepositions are referred to as stop-words; they have no inherent meaning and are therefore ignored in indexing and retrieval tasks. The removal of stop-words from Arabic text has a significant impact in terms of reducing the size of a cor- pus text, which leads to an improvement in the effectiveness and performance of Arabic-language processing systems. This study investigated the effectiveness of applying a stop-word lists elimination with normalization as a preprocessing step. The idea was to merge statistical method with the linguistic method to attain the best efficacy, and comparing the effects of this two-pronged approach in reducing corpus size for Ara- bic natural language processing systems. Three stop-word lists were considered: an Arabic Text Lookup Stop-list, Frequency- based Stop-list using Zipf's law, and Combined Stop-list. An experiment was conducted using a selected file from the Ara- bic Newswire data set. In the experiment, the size of the cor- pus was compared after removing the words contained in each list. The results showed that the best reduction in size was achieved by using the Combined Stop-list with normalization, with a word count reduction of 452930 and a compression rate of 30%.

Keywords:

Arabic, Normalization, Preprocessing, StopWords, Zipf's law 2012 ACM Computing Classification System: Computing methodologies, Artificial intelligence, Natural language processing, ACM Computing Classification System: Computing methodologies, Artificial intelligence, Natural language processing

1. Introduction

Natural Language Processing (NLP) is a branch of artificial in- telligence concerned with how to analyze and process human languages with the interactions in the environment of the com- puter. For example, NLP today makes it possible for machines to read text, interpret speech, hear it, gauge sentiment, and de- fine what the important parts of it.

Recently, research work on issues around the use of natu- ral language processing for the Arabic language has been con- ducted and good progress has been made. Nevertheless, this work and the efforts that have been made are still not suffi- cient when compared to the English language achievements that have been made in the NLP field, which is central to NLP research for a long time [1]. This comparative lack exists in spite of that the Arabic language in the United Nations is one of the six major languages as well as spoken by around 400 million people. Moreover, it is the second language of numer- ous people in the Islamic world because it is the language of the holy book Qur'an.

For a NLP system to perform well, it is essential that stop- words are removed from the text before it is processed. A stop- word is a commonly used word, such as a preposition, that has no inherent meaning and that therefore does not facilitate NLP. Hence a stop-word list that contains such words is usually used in a preprocessing step so as to remove them prior to

Manuscript received July 5, 2022
Manuscript revised July 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.7.9

processing

text. Stop-word lists can be developed in various ways. They can consist of words that are most frequently used or contain words that are considered unnecessary based on syntax, which needs the involvement of an expert who makes a decision to include or omit words based on a personal judgment [2]. In the specific case of Arabic NLP, stop-words could be translated from the English language into the Arabic language for inclu- sion in a stop-word list for an Arabic NLP system. A stop-word list could also be created by using a combination of syntactic rules, a frequency-based list of words in corpus and the translation of English-language stop-words into Arabic in order to gain the benefits of these approaches [3].

Crucially, there is no general standard stop-word list for NLP experiments for the Arabic language. One of the Arabic stop- word lists that has been used by some researchers [4] can be found in the Lemur Toolkit, containing 168 words, which is relatively short, and it was developed in [5], when she created a root-based Arabic stemmer. Table 1 show an example of in the Lemur toolkit.

This research try to present the effects of the elimination of stop-words with text normalization as a preprocessing step in reducing the size of an Arabic text corpus. For that, Three stop- word lists were considered: the abovementioned Lookup Stop- list of 168 stop-words [5], a Frequency-based Stop-list built by applying Zipf's law and based on the statistical distribution of words [6], and a Combined Stop-list containing the words from the previous two lists developed by us with the elimination of frequent words, which more details about lookup, Frequent, combined stop-words lists in section 4.3.

This paper is organized as follows: Section 2 highlights the research motivation. Section 3 gives some background to the study by providing an overview of related work. Section 4 de- scribes the experiment. Section 5 presents and evaluates the re- sults. Section 5 makes some concluding remarks and describes future work.

2. Motivation

The Arabic language is one of the most highly polysemy and complex morphological natural languages [7]; the word have more than one meaning depends on the context of the word. Such that "قلب" which means "heart" or "core". Re- searchers in [5] and [8] stated that in a large group of doc- uments, about 50% or less of the terms can be stop-words, which can affect the indexing process by deleting these terms. That motivate this work to enhance Arabic NLP system per- formance. However, NLP Arabic text systems still have many drawbacks in many of their processing steps such as prepro- cessing step. Hence the utilization of a stop-word elimination procedure and a normalization technique would have a posi- tive effect on Arabic NLP systems performance. Yet, there is no dominant stop-word list for the Arabic NLP system. More- over, there are no standard normalization steps. Therefore, the objective of this paper is to propose a method for developing stop-words list that enhance the NLP system in preprocessing step for Arabic-language texts in respect of one main aspect, namely, the efficient use of the stop-word list that is intended to improve the performance of the Arabic NLP framework.

3. Related Work

In the text, the skewed of the word frequencies distribution is one of the most apparent text properties from a statistical point of view. And for word distribution, many words have low recurrences while few have many recurrences. Actually, in English-language 40% of all English text is consist of the top 50 frequent words, while 20% of them for the six frequent words. Furthermore, in the English language “the” and “of” is the most two frequently words [6].

Stop-words carry little meaning but they are very common words that appear in text; they works only as syntactic task and do not refer to the subject matter. There are two different as- pects for stop-words effects on the NLP systems. Firstly, they can affect the system effectiveness because they appear in texts at very high frequency rate and tend to reduce the effect of frequency differences among less popular words, thereby affect- ing the weighting process. Moreover, the eliminations of stop- words can change the document length and still subsequently affects the weighting process. Secondly, stop-words can also affect the efficiency of system due to their nature and that are meaningless, which may costly and consuming large amount of time, yet unproductive processing according to [8] [9]. How- ever, some of the previous work such [6] and [10] mentioned that in a large corpus of documents around 50% or less of terms can be a stop-words, which can effects on the indexing process by eliminations of these stop words. Stop-words also called as “function words”, which is belong to different categories of words like conjunctions, prepositions, and adverbs as mentioned by [3].

Table 1: Examples of Arabic stop-words

Arabic Stop Word	English Meaning
ان	That
بعد	After
ضد	Against
يلي	Follows
■حتى	Until

Stop-words in all documents are the most frequent words, so they are count as weakly distinguishable [11]. In other words, the content of a text cannot be determined by depending on these words alone [6]. There are many benefits that can be gained from eliminating stop-words. Primarily, the resultant shortening of the indexing structure speeds up processing yet does not damage retrieval effectiveness [12] [13]. Additionally, the system is not burdened with unnecessary information [14]. The stop-word lists most often used for English-language NLP are rather small, consisting of 200 to 400 words, because of the multitude morphological of the language; the list consist of different morphological forms of each stop-word [15].

Despite all these benefits, there are some disadvantages as- sociated with eliminating stop-words such as an adverse effect on recall. Additionally, there are some sentences in the English language that are comprised solely of stop-words, for exam- ple, the famous Shakespearean sentence “To be or not to be”, which would be removed from the text if all stop-words were automatically eliminated.

The elimination of stop-words from the Arabic text is com- plex for its own reasons. Arabic is wealthy in vocabularies, which leads to the appearance of stop-words in a huge amount [16]. Furthermore, stop-words in Arabic have several specific properties that need to be considered [17]:

- They are meaningless when used alone.
- They be seen plenty in a text.
- They are mandatory for language construction.
- They consist generally of adjectives.
- They are public words and are not especially used in a specific domain.
- They are not important for any researcher.
- They can never shape a complete clause when utilized alone.

The BOW method and three levels of N-grams (3, 4, and 5) assessed the extraction, as results obtained by the Naïve Bayes classifier revealed that BOW outperformed all levels of N-gram tested [18]. Likewise, Researcher in [19] developed a method for searching for the Arabic text in retrieving Ara- bic information by applying two methods of information ex- traction, namely, contiguous N-grams and hybrid N-grams, the hybrid N-grams better than contiguous N-grams and both of which are Methods for extracting nouns and making confusion between nouns and words verbs when converting them to their roots "stem". Researchers in [18, 19] does not evaluate and respect the elimination of stop-words list in Arabic classifiers. Moreover, in [20] has stemmed nouns and verbs by using ETS as the stemmer for the roots of the Arabic words. The algorithm aims to select the noun and verb from Arabic documents according to prepositions, in addition to some rules related to other linguistic elements, such as the definition of the article "the". The algorithm can distinguish between nouns and verbs

but nouns are abstracted and have limited extraction rules for verbs. They generates initiation stopping words of nouns and verbs without assessing their effect on Arabic clustering, it is the main objective of current work to assess the effect of delet- ing stop words on Arabic NLP systems.

Arabic stop-words are fundamental as they contain the def- inite article (Al; the) as grammatical links, joined and de- tach preposition, conjugation(s), questioning words, negative words, interjection and proclamation letters, the juncture of place and time, in addition to all consciences, expository, ob- ject and subject consciences, discriminatory Nouns, some num- bers, additions, and verbs. Also, stop-words may be detached or joined in the shape of suffix as mentioned by light [17] or prefix as stated by Khoja [21]. Hence their removal presents the NLP researcher with additional challenges.

4. Experiment

This study scout the effect of stop-words and their utilization on reducing the size of the Arabic text corpus. It tries to employ different scenarios for removing stop-words from the corpus by comparing the resultant effect on corpus size. In this study, three stop-lists were considered: an Arabic Text Lookup Stop- list, Frequency-based Stop-list using Zipf's law, and Combined Stop-list. These stop-lists were applied to a 1733 files from the Arabic Newswire data set [22].

The study evaluated the application of the stop-word list technique by using the number of eliminated terms and the number of terms in the corpus after elimination as the basis for comparison. First, the elimination of the stop-words in the Lookup Stop-list without normalization was checked. Then, a different scenarios in comparison between the elimination of the top frequent terms in the corpus and the elimination of the Lookup Stop-list was conducted. Finally, the effectiveness of using a combination approach consisting of the Lookup Stop- list and the top frequent terms in the corpus were tested. Fur- thermore, because text normalization step(s) guarantee a fixed order of characters where multiple variants are allowed, each stop-list was tested before and after text normalization. The effectiveness of all the scenarios for removing stop-words lists with or without normalization evaluated to define which stop- list was able to achieve the optimal compression rate for an Arabic text corpus.

4.1 Data set

This research used 1733 documents of an Arabic test corpus that was created by the LDC in Philadelphia, U.S.A., which was used in TREC (The Text REtrieval Conference) experi- ments, and which was developed by David Graff and Kevin Walker at the LDC [22].

4.2 Normalization

In NLP experiments, the normalization of data is important be- cause texts such as newspapers and documents may use variant encoding rules (or no rules at all), even for the same language. Many inconsistencies that result in preprocessing (e.g., unreal- ized stop-words) or in text-matching (e.g., spelling variation in the article text) could be appeared due to the mistakes in text normalization.

This issue is further compounded in the Arabic language case, where to refer to short vowels, words could be formed by

diacritics or without it. Such that, عِلم and علم (u'lem and alam) both for eyes look like, but in the computer, they do not look same. Short vowels are generally not included in texts, such as newspapers, which account for a vast propor- tion of the texts that are searched online. For this reason, some types of normalization, such as the removal of diacritics, punc- tuation, and non-letters are typically performed in NLP sys- tems.

In this experiment, the documents was normalized by fol- lowing the same normalization steps used by [23] [4], as fol- lows:

- Eliminate punctuation
- Eliminate pronunciation signs (mainly slight

vowels), the eliminations of which makes text harmonic.

- Eliminate number and special characters.
- Replacement of أ ,آ, and إ with ا
- Replacement of final ى with ي
- Replacement of final ة with ه

4.3 Stop-lists

Three stop lists were created. The first one, the lookup list which consist of 168 words, the same list created by Khoja and used in the lemur toolkit. The second one, a list of 135 words the most frequent words created based on the distribution of the words using Zipf's law. Finally, the third list, a combined list consists of both previous lists by removing the duplication of words. As mentioned before, one of the obvious facts when dealing with texts that the word frequencies distribution is very swerve from a statistical point of view. There are many words that have few frequencies and there are few words that have numerous frequencies. In fact, For the English language ("the" and "of") they are the most two frequent words that appear in any text. The 20% of appearance is for the topmost six recur- rent terms, and 40% for the top 50 frequent words. On the other side, for a large corpus of text, the typical frequency is only one is the frequency for half of the unique words in that corpus. This is represented by Zipf's distribution law, which clarifies the recurrence (f) of the word times its rank is roughly static (k), or as an alternative, the recurrence of the r^{th} most popular terms is oppositely symmetrical to r is calculated us- ing equation (1):

$$k = f.r \tag{1}$$

Thus the probability of a word occurring in a text is calculated by the recurrence of the term divided on the summation of term frequencies in the corpus. So, therefore, Zipf's law is repre- sented in equation (2) as:

$$c = r.Pr \tag{2}$$

where Pr is the likelihood of appearance for the rth ranked word, and c is fixed. For English text, c ≈ 0.1 [6].

In this study, a Frequency-based Stop-list was created based on Zipf's law from the most frequent terms in the selected data set. In general, this list contained variant syntactic categories in Arabic without any formal method applied to guarantee the per- fection of the list. Furthermore, this list contained a few words that appeared in the Lookup Stop-list. The word categories that were collated for the Frequency-based Stop-list were as follows:

- Adverbs
- Pronouns
- Prepositions
- Nouns
- Other.

The list created from the top 135 terms appeared as the most frequent words, as some of them were not considered stop- words before. Table **2** shows the top five most frequent words in the data set.

The second stop-word list used in this study was the Lookup Stop-list that contains 168 stop-words. This list was developed by [5] and has been used by many other researchers includ- ing [24] [4] The reason for the small quantity of stop-words in the Lookup Stop-list is because the properties of the words themselves, where the most frequent terms in the Arabic lan- guage mainly consist of prepositions, pronouns, and adverbs. **Table 3** provides a sample of some of these words.

Finally, the third stop-list that was considered in this study was a Combined Stop-list which was comprised of both lists (i.e. lookup list and most frequent words list) that were com- bined together by removing the duplicated words. In contrast to [10], because of the properties of Arabic there is a small number of stop-words. Firstly, in Arabic, many characters could be utilized as prefixes and may convert the meaning of a term. These characters are ("ف" ,"ل" ,"ك" ,"ب","أ"), they were employed on several of the stop-words as a prefix, that they could be used as part of the original text, removing them could removing important words . Secondly, a large num- ber of the main words in the documents as stop-words that can be combined with one another and make use as affixes (i.e. suf- fixes or prefixes), especially pronouns, but these in general are rarely used and could change the meaning of a word and reflect an important meaning in the context of an individual document. Finally, the conjunction letter "waw" meaning "and" could be used as a suffix for a word and could be a part of the original word. Hence, the elimination of such letter is often subject to a

Table 2: Top five most frequent Arabic stop-words using Zipf's law.

Word Rank	Arabic Word	Zipf Frequency	English Meaning
1	نادي	1483.00	Club
2	غزة	741.50	Gaza
3	خلال	494.33	Through
4	وقال	370.75	He said
5	العام	296.60	The year

Table 3: Sample of stop-words in the Lookup Stop-list.

Arabic Word	English Meaning	Arabic Word	English Meaning
في	in	التي	which
الا	except	من	from
اول	first	الى	to
انه	It's a	على	on
ال	The	بعد	after

certain decision. So that, there are no obvious steps to develop of a stop-words list for Arabic.

In this study, the following steps were followed to test the three stop-lists:

1. Remove the stop-words in the Lookup Stop-list from the data set

2. Remove the stop-words from the Frequent Stop-list from the data set

3. Remove the words from both the Lookup Stop-list and Fre- quent Stop-list from the data set.

4. Calculate the size of the data set and the rate of compres- sion (i.e., the reduction rate) of the text after each step.

5. Check the effect of normalization on stop-word elimina- tion by performing steps 1 to 4 before and after normalization.

It is unclear what words could be treated as stopping words and which could not. The conventional procedures deal with the terms that are frequently appeared multiple times are stop- ping words, However, several recurring terms in a given text that are deemed essential terms such as "indexing terms". How- ever, while a text topic is more specialized, the use of frequent terms, So that the indexing terms become useless as indexing terms [25].

without normalized stop-words and with text normalization, and named as (DSWTNSWNText).

4. The data set size was calculated as the number of

So as to assess the efficacy of removing stop-words from the text as described in the abovementioned steps, some compar- isons were made to ascertain how the elimination of stop-list affected the volume of the resulting corpus. To do this, the size of the data set was measured a number of times: before delet- ing the frequent terms; after removing the stop-words in the Lookup Stop-list only; after removing the frequent terms only; and after removing both the frequent terms and the stop-words in the Lookup Stop-list.

8. Result and Evaluation

This study compared the effect of using three stop-word lists on an Arabic text corpus in terms of the resulting size of the cor- pus and the compression rate. Different steps were applied in the experiment to determine which of the lists with or without normalization was capable of achieving the best results. The steps that were applied in this experiment are detailed below:

1. The data set size was calculated to determine the total number of words, and named as (DS).

2. The data set size was calculated as the number of words without normalized stop-words and without text normalization, and named as (DSWTSWWTN).

3. The data set size was calculated as the number of words

words with normalized stop-words and without text normalization, and named as (DTWNSWWTNText).

5. The data set size was calculated as the number of words with normalized stop-words and with text

normalization, and named as (DSWNSWNText).

6. The data set size was calculated as the number of words without normalized most frequent words and without text nor- malization, and named as (DSWMFW).

7. The data set size was calculated as the number of words without normalized most frequent words and without text nor- malization, and named as (DSNTextWTMFW).

8. The data set size was calculated as the number of words with normalized most frequent words and with text normaliza- tion, and named as (DSNTEXTWNMFNSW).

After each one of these steps the data set size was measured in terms of the number of words and the compression rate. These results are presented in Table 4.

Figure 1 shows the size of the corpus after elimination of each set of stop-words in terms of total number of words re- maining after each step in the experiment. **Figure 2** shows a comparison of the compression rates after the removal each set of stop-words after each step in the experiment. This work contributes to extracting the most important terms by control and reduce the rate of non-important terms called stop-word list. The findings could help the pre-processing process of the natural language processing domain to exert effort as well as time on the detection of the most important terms such as reduce the number of big features. This research is limited to modern Arabic language.

9. Conclusion and Future Work

This paper examined the efficacy of using three variant stop- words lists in preprocessing of Arabic text for NLP. The three lists in question were the Lookup Stop-list, which was taken from the literature; a corpus-based stop-word list, which con- sisted of the most frequent words in the corpus which were identified by applying Zipf's law; and a Combined Stop-list, which included the words contained in both of the preceding lists. These three stop-word lists were tested using different scenarios with and without normalization. As a stop-word list is often defined as a list containing the most frequent words that do not have meaning in themselves, and there is no stan- dard stop-list for most of the natural languages, the researcher used stop-words consisting of prepositions adverbs, and pro- nouns. The results showed that removing the most frequent corpus words as well as those in the Lookup Stop-list with nor- malization was the best way in which to reduce the corpus in terms of number of words and compression rate.

It can therefore be concluded that using a stop-word list de- veloped in the same way as described and following the tech- nique used in this study will reduce the size of a data set or a corpus and enhance the compression rate. Furthermore, this technique takes into consideration the stop-words that are of relevance to the particular domain under study. In general, it is believed that this technique could be applied to retrieval sys- tems or to automated text mining. To assess the performance of this technique, extensive experimental studies need to be con- ducted to examine retrieval performance or mining accuracy especially for Arabic text, which will be a future work.

Acknowledgment

We would like to thank the Linguistic Data Consortium (LDC) for providing us with the LDC2001T55 Arabic Newswire Part 1 data set at no cost, and for awarding us with the fall 2012 LDC Data Scholarship.

References

[1] R. Elbarougy, G. Behery, and A. Khatib, “A Proposed Natural Language Processing Preprocessing Procedures for Enhancing Arabic Text Summarization,” Studies in Computational Intelligence, vol. 874, pp. 39–57, 2020.

[2] C. Fox, “A stop list for general text,” Acm Sigir Forum, vol. 24, no. 1-2, pp. 19–21, 1989.

Table 4: Data set size after each experiment with compression rate.

Arabic Word	English Meaning	Arabic Word
DS	646745	1.000
DSWTSWWTN	501884	0.776
DSWTNSWNText	501573	0.776
DTWNSWWTNText	533478	0.825
DSWNSWNText	497939	0.770
DSWMFW	474247	0.733
DSNTextWTMFW	470713	0.728
DSNTEXTWNMFWNSW	452930	0.700

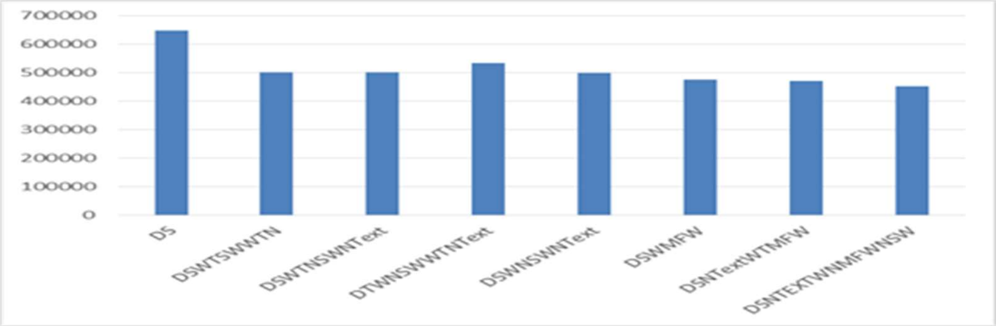

Fig. 1 Difference in data set size after each experimental scenario.

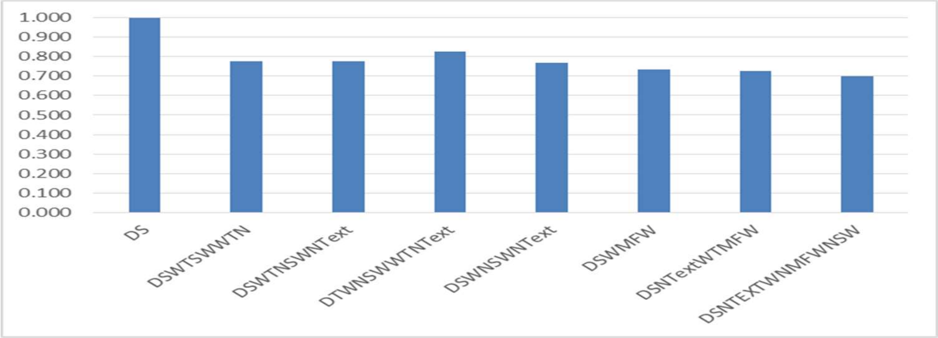

Fig. 2 Difference in compression rate after each experimental scenario.

[3] E. T. Al-Shammari, “Lemmatizing, stemming, and query expansion method and system,” Google Patents. Available at, 2013.

[4] L. S. Larkey, L. Ballesteros, and M. E. Connell, “Light stemming for Arabic information retrieval,” Arabic computational morphology, pp. 221–243, 2007.

[5] S. Khoja, “APT: Arabic part-of-speech tagger,” Proceedings of the Student Workshop at NAACL, pp. 20–25, 2001.

[6] W. B. Croft, D. Metzler, and T. Strohman, “Addison-Wesley Reading,” Search engines: Information retrieval in practice, vol. 520, 2010.

[7] Al-Shalabi, Riyad, G. Kanaan, M. Yaseen, B. Al- Sarayreh, and N. Al-Naji, “Arabic query expansion using interactive word sense disambiguation,” in *Proceedings of the Second International Conference on Arabic Language Resources and Tools*, 2009.

[8] A. Masrai and J. Milton, “How different is Arabic from other languages? The relationship between word frequency and lexical coverage,” Journal of Applied Linguistics and Language Research, vol. 3, no. 1, pp. 15–35, 2016.

[9] Y. Hacohen-Kerner, D. Miller, and Y. Yigal, “The influence of preprocessing on text classification using a bag-of-words representation,” PloS One, vol. 15, no. 5, 2020.

[10] I. A. El-Khair, “Effects of stop words elimination for Arabic information retrieval: a comparative study,” International Journal of Computing & Information Sciences, vol. 4, no. 3, pp. 119–133, 2006.

[11] S. Sarica and J. Luo, 2020.

[12] A. W. Pradana and M. Hayaty, “The effect of stemming and removal of stopwords on the accuracy of sentiment analysis on indonesian-language texts,” Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control, pp. 375–380, 2019.

[13] E. L. Lydia, P. K. Kumar, K. Shankar, S. K. Lakshmanaprabu, R. M. Vidhyavathi, and A. Maseleno, “Charismatic document clustering through novel K-Means non-negative matrix factorization (KNMF) algorithm using key phrase extraction,” International Journal of Parallel Programming, vol. 48, no. 3, pp. 496–514, 2020.

[14] R. Baeza-Yates and B. Ribeiro-Neto, 1999.

[15] Al-Shalabi, Riyadh, G. Kanaan, J. M. Jaam, A. Hasnah, and E. Hilat, “Stop-word removal algorithm for Arabic language,” Proceedings. 2004 International Conference on Information and Communication Technologies: From Theory to Applications, vol. 545, 2004.

[16] H. Schütze, C. D. Manning, and P. Raghavan, 2008.

[17] B. Alhadidi and M. Alwedyan, “Hybrid Stop-Word Removal Technique for Arabic Language,” Egyptian Computer Science Journal, vol. 30, no. 1, pp. 35–38, 2008.

[18] B. Al-Salemi and M. J. A. Aziz, “Statistical bayesian learning for automatic arabic text categorization,” Journal of Computer Science, vol. 7, no. 1, 2011.

[19] S. H. Mustafa, “Character contiguity in N-gram-based word matching: the case for Arabic text searching,” Information Processing & Management, vol. 41, no. 4, pp. 819–827, 2005.

[20] E. Al-Shammari and J. Lin, “A novel Arabic lemmatization algorithm,” Proceedings of the Second Workshop on Analytics for Noisy Unstructured Text Data, pp. 113–118, 2008.

[21] J. Atwan and M. Mohd, “Arabic Query Expansion: A Review,” Asian Journal of Information Technology, vol. 16, no. 10, pp. 754–770, 2017.

[22] A. Cole, D. Graff, and K. Walker, “Arabic Newswire Part 1 Corpus (1-58563-190-6),” Linguistic Data Consortium (LDC). Available at, 2001.

[23] B. F. Willian and B. Y. Ricardo, 1999.

[24] J. Atwan, M. Mohd, H. Rashaideh, and G. Kanaan, 1999“Se- mantically enhanced pseudo relevance feedback for ara- bic information retrieval,” Journal of Information Sci- ence, vol. 42, no. 2, pp. 246–260, 2016.

[25] J. Atwan, M. Mohd, and G. Kanaan, “Enhanced arabic information retrieval: Light stemming and stop words,” International Multi-Conference on Artificial Intelligence Technology, pp. 219–228,