Sana'a University Journal of Applied Sciences and Technology https://journals.su.edu.ye/index.php/jast
Vol. 1 | No. 4 | Page 313 – 326 | 2023 | ISSN: 2958-9568
Doi: 10.59628/jast.v1i4.588

Comparative Analysis and Evaluation of Stemming and Preprocessing Techniques for Arabic Text

Abdualmajed A. G. Al-Khulaidi[1, *], Samer Mohammed Yaseen[1, *]
[1] Department of Computer Science, Faculty of Computer and Information Technology, Sana'a University, Sana'a, Yemen.
***Corresponding author:** alkhulaidi@su.edu.ye and sameer@su.edu.ye*

ARTICLE INFO
Article history:
Received: Aug 27, 2023
Accepted: Oct 29, 2023
Published: Nov, 2023

KEYWORDS
1. Natural Language Processing
2. Information Retrieval
3. Arabic Information Retrieval
4. Stemming
5. Text Preprocessing

ABSTRACT
Arabic information retrieval is challenging due to the language's complex morphology and syntax. Preprocessing and stemming improve the accuracy and efficiency of Arabic information retrieval. This paper provides a comprehensive analysis of the existing literature on Arabic preprocessing and stemming techniques. The paper identifies the limitations and challenges of these techniques. The paper emphasizes the importance of preprocessing and stemming and underscores the need for further research to improve Arabic information retrieval. This study evaluates ten stemmers on a public dataset. The results show that root-based stemmers: Lucene, and khoja got the highest reduction rate 90.9%, and 85% respectively. The results emphasize that root-based stemmers have good conflating ability for similar terms, while light-based stemmers under-stem similar terms.

CONTENTS

1. Introduction
2. Related Work
3. Methodology
4. Experiments and Results
5. Discussion
6. Conclusion
7. References

1. Introduction:

With the growth of shared information in recent decades, there has been a need for implementing mechanisms to access specific information. Information retrieval (IR) systems are now among these mechanisms. IR is a field of research that focuses on finding information from unstructured sources, such as text, which satisfies information needs, such as user queries[1]. Standard IR consists of four main phases: preprocessing, indexing, querying, and IR[2]. The IR system accepts the user query, which is a set of words describing the user's needs, and returns some documents ranked according to the similarity between the user query and the documents' content. The first key step in an IR system is stemming [3]. Many factors contribute to IR performance. One of the most factors that affect the performance is the stemming [4]. Stemming is the process of removing the affixes of words to group several related terms into one index term [5] ; [2]. For example, the terms "كتب"*ktb*. "يكتب"*yktb*, and "اكتب"*aktb* can be conflated to the term "كتب"*ktb* which hold the same concept of the previous three terms. By indexing the documents using the stem of the word, several words will be conflated to one stem, which in turn will save the space required for IR system and enhance the computational load of the system [2]. Stemming has a high impact on Arabic text retrieval [6].

However, the importance of stemming in Arabic IR, it still an open issue and need improvement [7]. Therefore, this paper illustrated the stemming approaches used in Arabic IR. We will analyze the approaches and discuss the strengths and weaknesses of each. This paper is organized as follows: The first section introduces the topic. The second section explores the main characteristics of the Arabic language. The third section highlights the significance of preprocessing and stop words removal in Arabic IR. The fourth section delves into the various stemming approaches used in the Arabic language. Finally, in sections 3, and 4, we examine some stemming techniques and report their effectiveness in conflating similar Arabic terms.

1.2 Arabic Language Characteristics

The Arabic language is written from right to left and has twenty-eight characters. The letters of Arabic language are connected to form words. Each letter has different shapes depending on its position in the word, whether it is at the beginning, middle, end, or separated. Arabic language is ranked as the fastest-growing language in the number of internet users [8]; ([9]. The Arabic language is rich in vocabulary, a single root composed of three to five letters can generate enormous number of words with different meaning [10]. According to [8], Arabic has an estimated of sixty billion words derived from about 10,000 roots, making it a highly derivational language [11]. For example, the Arabic root "**علم**" can generate many words, such as "**تعليم**", "**عالم**", "**معلم**", and "**تعلم**". These words can also accept prefixes and suffixes, adding more words, such as "**معلمة**", "**تعليمهم**", "**تعلمها**", "**تعاليم**", "**تعليمات**", "**متعلم**", and "**متعالم**".

2 Literature Review

Preprocessing and stemming in Arabic information retrieval have been extensively studied by researchers. This section aims to provide a comprehensive review of significant research conducted on addressing these challenges. The review will begin by examining various preprocessing techniques employed in Arabic information retrieval, followed by an exploration of different approaches to stemming.

2.1 Preprocessing and Text Normalization

The preprocessing and normalization of text are essential steps in stemming for any language. The normalization step involves removing some letters such as "**و**" or removing the diacritics and "*hamzah*". The preprocessing step involves removing stop words such as "and", "or", "the", "for", etc. In Arabic, similar words such as **"ال"**, **"من"**, **"الى"**, **"عن"**, and **"على"** are removed. This process improves the performance and storage capacity according to [2]; [8] ; [12]; [7];[13]. Removing stop words also improve accuracy [14]. The effect of stop word list removal in Arabic IR was studied by [15]. He evaluated three stop word list and reported that stop word list removal improved retrieval effectiveness, especially when used with the BM25 weight. The length of documents can influence the effectiveness of stop words removal, and satisfactory results might appear when removing stop words from long documents according to [8]. However, the importance of stop word removal, there is no standard stop word list in Arabic IR [2]. May be this is because Arabic stop words can be combined with prefixes or suffixes, such as the word **"عند"** which can appear as**"عندهم"** , or **" عنده "** Which represent another challenge [8]. Therefore, to improve Arabic IR, stop word list need to include all affix possibilities, and to include stop words used in different Arab countries[11]. [16] investigated and found that the best results they got are obtained by combining Khoja stop word list with [17] stop word list. To summary, stop word removal is a crucial step in Arabic IR. However, a standard stop word list is not available. Researchers have suggested various stop word lists, and the effectiveness of these lists varies based on the dataset used and the techniques applied. Therefore, selecting an appropriate stop word list is essential to achieve optimal performance in Arabic IR.

2.2 Stemming techniques

There are two types of stemming in the literature. The first type is a rule-based approach

that involves removing the affixes of the word based on several word shape rules, but no linguistic rules are considered. In this approach, the resulting stem may not have a proper meaning, and the stem may not be an actual word. The second approach is the linguistic approach, which involves analyzing the grammatical structure and meaning of a word to determine its stem. This approach is based on linguistic theories and models that consider the syntactic and semantic relationships between words in a language. For example, a linguistic-based stemmer might use part-of-speech tagging and dependency parsing to identify the stem of a word based on its role in a sentence. Linguistic-based stemming tends to be more accurate than rule-based stemming. but can be computationally expensive and may require more linguistic knowledge and resources. In this research, we will focus on the rule based stemming techniques as they are used widely in IR tasks. The purpose of rule-based stemming techniques is to group similar terms together, and not necessary to find the accurate root of the word [18]. Rule-based stemming is divided into two approaches: root-based and light-based stemming [19];[20]. The early experiments of Arabic IR relied on a small collection of data [8], and the stemming task was done manually by an expert in the language [21]. [4] is an example of such manually stemming approaches. However, manually stemming for an IR task is not efficient [21]. So, in this section, we will introduce the algorithmic approaches to stemming.

2.2.1 Root based stemming.

In this section, we will introduce root-based stemming techniques. Khoja stemmer [22] is one of the most famous examples of root-based approaches. Khoja stemmer removes suffixes, prefixes, apply set of rules and pattern match based on word length to adjust the stem, and at the end it compares the result stem against a list of roots. [23] proposed a root-based stemmer that tries to extract the root of the word. This stemmer has three modules. The first module is the build module, which uses ALPNET morphological analyzer to generate a list of word-root pairs. The second module estimates the possible prefixes and suffixes of Arabic words. The third module is the detect module, which is used in action to eliminate the possible prefixes and suffixes of input word. [18] propose a root-based stemmer called the ISRI stemmer, which is like Khoja stemmer but without using a root list. This stemmer starts by checking the word against several patterns based on word length. When no match is found, the stemmer then removes specific prefixes and suffixes based on word length and tries again to find a pattern match to extract the root. Another root-based stemmer is proposed by [24]. Their stemmer is based on 15 steps. The first step in this stemmer is to check the word length. If it is greater than three, then the rest of the steps are applied. If not, the word might be a root or a stop word. After doing all the modifications in Table 1, the word is checked against several patterns like the Khoja stemmer with a slight difference. [25] use a modified version of the Ghwanmeh stemmer to extract three indexing terms: stem, verbed pattern, and root. For each word in a document, the three terms are extracted and indexed with their frequencies.

Table 1: some of Gwanmah stemmer actions

Action	Letter	Place of Action	Condition
Remove	ال	prefix	word > 3
Replace	إ with ا	All the word	word > 3
Remove	كم، كن، تم، تن، ين، ان، ين، ات، ون، ها، ية، هم، نا ، ما، وا، ني، يا ، هن	suffix	word >=6
Remove	كال، بال، فال، مال، وال،	prefix	word >=6
Remove	سن، سي، ست	prefix	word >4
Remove	ل	Prefix	word >3
Remove	ه، ة، ت	Suffix	word >3
Remove	ت، ي	Prefix	word >3
Remove	ا	Suffix	word >=4

The computation time and the required space capacity are some of the concerns of this approach. [26] improved the ISRI stemmer by adding a rule for words of length two. They

checked if these words have a weak letter. They evaluated the improved version on the Quranic Arabic Corpus and showed that the improved version can detect the root of length two correctly compared to the original ISRI. [27] also proposed a modified version of the ISRI stemmer by adding several rules to face broken plurals problem. They evaluated their modified stemmer on the TREC dataset and showed that their stemmer achieved remarkable results compared to ALSR2 and the original ISRI stemmers.

2.2.2 Light Based Stemming

Light-based stemming is a simple approach that involves removing certain prefixes or suffixes from words [11]. [6] proposed several light stemmers, called light 1, 2, 3, and 8, which remove specific prefixes or suffixes from words as shown in Table 2. [28] modified the stemmer of [29] by adding a word normalization step that removes stop words and punctuation marks.

Table 2: larkey et.al prefixes and suffixes list.

Stemmer	Prefixes	suffixes
Light1	بال، وال، ال، فال آال،	**None**
Light 2	بال، وال، ال، و فال، آال،	**None**
Light 3		ة ه،
Light 8		ين، ون، ات، ان، ها، ي ة، ه، يه، ية،

[30] proposed a new stemming technique that removes several infixes and searches for equivalent results in a lexicon. [21] introduced a new light stemmer called light10, which removes the letter "**و**" and the definite article "**ال**" from the beginning of words and some suffixes.

[21] concluded that even though their approach is simple, it shows highly effective results compared to the root-based stemmers. [3] proposed an enhanced stemmer that tries to handle the problem of broken plurals by using two prefix lists, two suffix sets, and a validation step. [31] proposed the SAFAR stemmer, which generates several stems for each word and evaluates the results on a lexicon. [32] mixed the Khoja and light-based stemmers by using Khoja stemmer for verbs and light10 stemmer with an extended prefix and suffix list for nouns. [5] proposed the AMIR stemmer, which can remove infixes along with prefixes and suffixes to extract the stem of the word. Finally, [33] proposed a stemmer called D-light stemmer, which have a list for definite articles, suffixes, and prefixes. This stemmer applies several rules based on word length to remove the definite article and suffixes before removing prefixes as shown in Table 3. Overall, light-based stemming techniques are simple and, but they have limitations in handling complex morphological features such as broken plurals and infixes.

Table 3: Al-shalabi et.al D-light stemmer rules

Condition	Definition Article Letters to be removed
Word >=7	وبال لبال، فبال،
Word >=6	الا ولل، كال، وال، فال، بال،
Word >=5	لي اا، ال، ، لل،
Word >=4	ل
Suffixes	**Letters to be removed**
Word >=8	كموها ناهما، ناكمو،
Word >=7	,'اعنا' ,'اعكم' ,'ناكم','موهن' ,'موهم' ,',' ,'ناهم' ,'انهم' ,'ونهم' ,'اوهم' ,'نوهن' ,'اعهم' ياته,'اتهم' ,'اتها' ,'توهن' ,'اننا' ,'ونهم' ,'ونكم'
Word >=6	,وها ,هما ,ناه ,اءه ,ونه ناك، تان، نها، تنا، , تان , هات , نكم , تهم , تها , وني وهن , وهم ,نهم يهم , انك, انك , ونك , ونه , وكم , تهن ,
Word >=5	ين , ان , ات , ون , وا , تا تم، نا، ون، ين، ان، ما , وا , ني , كن , تم , ها , يا ,نا , هن , كم , تن , وه , وت , وك, هم , ها , ان , هم ,
Word >=4	ن ا، ك، ت، ي، ه، ة،
Prefixes	**Letters to be removed**
Word >=8	افاست' ,'لاست' ،' فليست' ،'وليست
Word >=7	,والم ,انهم ,واست ,فاست ,ويست ,'أتست , فاست ,والا ,الاس كمست ,باست
Word >=6	تست، يست، فان، فلل، فلن، فلي، ولت، مست، بمس ,وست , فلأ , وسن , فلي وسي , است,

2.3 Root based VS Light based.

The pros and cons of light based, and root-based stemming techniques are discussed in literature. [34]is one of the early research projects stated that using the root and light based is better than using the word in retrieval. They also concluded in their experiment that both roots based and light-based work well, but the root-based work better in high recall levels. [25]; [19] mentioned the same that using light based in indexing result in better precision, while the root based perform better in recall. [35]

evaluated three stemming methods and suggested using hybrid method for Arabic language. [16] use the light10 stemmer in their research and concluded that light stemmer outperforms ISRI and ETS stemmers. The discussion continues to involve characteristics of better stemmer. [21] mentioned two types of errors in stemming, the weak stemmers fail to conflate related terms that should be grouped together, while strong stemmers conflate larger stems in which unrelated terms erroneously conflated. [36];[3]; [20]; [37]mentioned the main characteristics of using root-based approach and light-based approach. They stated that using root-based approach led to an over-stemming problem, which is grouping unrelated terms to single root. In the other hand using the light-based causes under-stemming problem which is the inability to group related terms to one stem. [38] investigate this case using Lucene open-source IR with light10 and Alkhalil morphological analyzer and found that the performance is differ based on the size of the query. For short queries, the root based got better results. For long queries, the light-based wins. [7] investigated the impact of preprocessing techniques on Arabic classifications and concluded that stemming still an open issue and need improvement.

3. Methodology

To investigate the issue of under stemming and over-stemming problem mentioned in the literature, we conducted an evaluation test for several stemmers. The test is made for root based and light-based stemmers. The tested stemmers are Khoja, Tashphyne, Assem, Cog, ISRI, Lucene, Farasa, light8, light10, and D-light.

3.1 Dataset preparation

The dataset used in this test is a collected dataset of similar terms distributed in thirty-four classes. The dataset term distribution is shown in (figure 1). The dataset is available online[1] in the following reference [39].

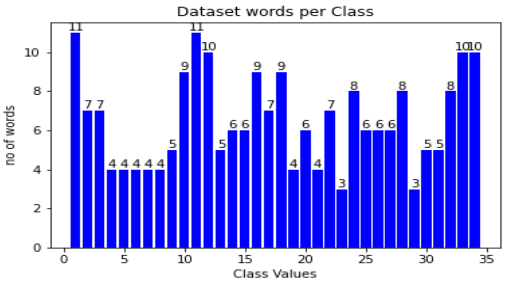

Figure. 1: Dataset word distribution

4. Experiments and Results

Two tests were conducted. The first test to show the ability of stemmers in stemming all dataset words. The second test conducted to show the ability of each stemmer in conflating related terms.

4.1 Test Environment

The test is implemented using python programming language. Some stemmers were available online, so we evaluate them against the dataset and report their results. Other stemmers were not available, so we implemented them like light8, light10 and D-light stemmers. A Jupiter notebook which includes the code and results for this implementation can be found here[2].

4.2 Stemming Ability

4.3 All stemmers, except khoja stemmer, were able to successfully stem all classes in the dataset. However, khoja stemmer showed incomplete stemming for class numbers 17 and 18 as shown in (figure 2).

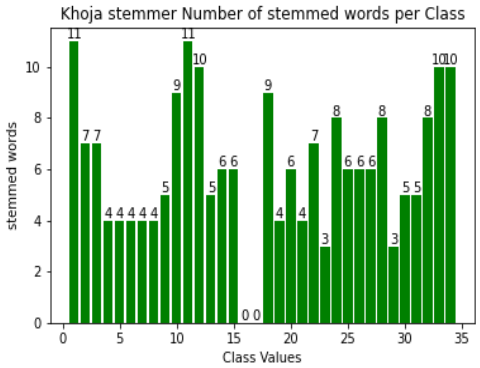

Figure. 2: Stemmed classes of Khoja stemmer

[1] The dataset can be accessed on the following link

[2] A jupyter notebook which has the code and details for this test can be found online in the following link

4.4 Conflating Ability

Given the dataset's consistent of similar terms within each class, it is crucial for stemmers to group each class into the smallest possible number of unique stems. as shown in (figure 3, and 4) lucene and khoja stemmers are the best in conflating similar terms.

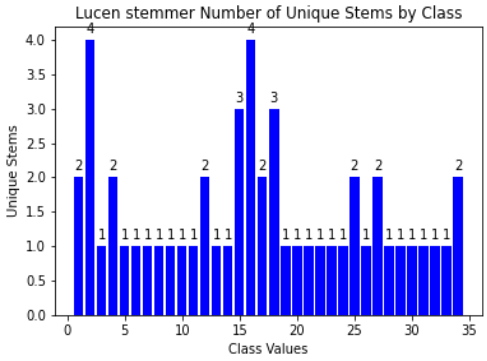

Figure. 3: Lucene stemmer unique stem result per class

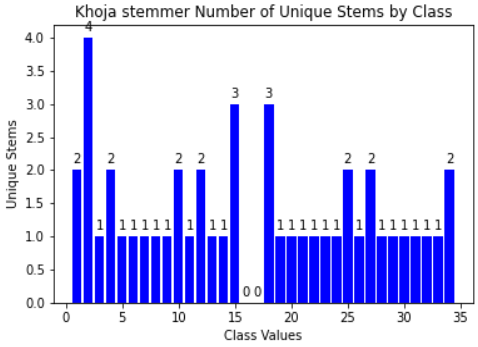

Figure. 4: khoja stemmer unique stem result per class

Additionally, it is worth noting that some stemmers have conflating ability better than others. As an example, ISRI stemmer has conflating ability better than the other stemmers (Cog, Farasa, Tashaphyne, light8, light10, D-light, and Assem). see Table 4, which shows comparison per class for all stemmers. The comparison is to show the ability of each stemmer in conflating similar terms to small no of unique stems. in general, light-based stemmers have less conflating ability compared to root-based stemmers.

4.5 Evaluation criterion

To evaluate the precision of stemmers, we employed the following formula*: **(Number of stemmed words, S – number of unique stems, U) / optimal value, O***. The optimal value represents the ideal scenario where each class is reduced to a single root, resulting in a total of thirty-four roots for the entire dataset. We calculated the optimal value by subtracting the number of optimal roots (34) from the total number of words in the dataset, which equated to **187 (221-34=187).**

4.6 Experimental results

Applying the previous formula to the khoja stemmer, we obtained a reduction rate of **(205 - 46) / 187 = 0.850**. Similarly, for the lucene stemmer, the reduction rate was **(221 - 51) / 187 = 0.909**. These results highlight the ability of lucene and khoja stemmers in conflating similar terms effectively. Table 5 summarizes the results for stemmed words, unique stems, and reduction rate produced by all stemmers.

Table 4: Stemmer's comparison of conflating ability per class

class	Dataset	Assem stemmer	Cog stemmer	Farasa stemmer	Lucene	ISRI stemmer	Khoja stemmer	Tashaphyne stemmer	Light8 stemmer	Light 10	D-light
1	11	11	11	6	2	4	2	6	11	11	11
2	7	5	5	5	4	4	4	5	5	7	5
3	7	6	6	5	1	3	1	6	7	6	6
4	4	3	3	4	2	2	2	4	3	4	3
5	4	3	4	3	1	2	1	4	4	4	4
6	4	3	2	3	1	3	1	3	2	4	2
7	4	4	3	4	1	3	1	3	3	4	3
8	4	4	4	4	1	3	1	3	4	4	4
9	5	4	4	3	1	1	1	3	4	5	4
10	9	7	8	6	1	3	2	4	8	9	7
11	11	8	9	7	1	2	1	7	9	11	8
12	10	8	9	8	2	6	2	7	9	10	8

13	5	5	5	4	1	2	1	4	5	5	5
14	6	4	4	5	1	3	1	4	6	6	5
15	6	6	6	6	3	4	3	6	6	6	6
16	9	7	6	7	4	5	0	4	7	9	6
17	7	7	7	5	2	7	0	7	7	7	7
18	9	9	9	7	3	7	3	9	9	9	9
19	4	4	4	4	1	1	1	4	4	4	4
20	6	5	6	6	1	3	1	5	6	6	4
21	4	3	2	2	1	1	1	2	2	4	2
22	7	5	6	6	1	1	1	6	6	7	6
23	3	3	3	3	1	1	1	3	3	3	3
24	8	4	4	5	1	1	1	4	4	8	4
25	6	5	4	5	2	4	2	5	5	6	4
26	6	6	6	5	1	1	1	6	6	6	6
27	6	5	4	4	2	2	2	5	4	6	4
28	8	8	8	6	1	2	1	8	8	8	8
29	3	3	3	3	1	2	1	3	3	3	3
30	5	5	5	4	1	1	1	5	5	5	5
31	5	4	4	3	1	1	1	3	4	5	4
32	8	5	5	5	1	7	1	6	5	8	6
33	10	9	9	10	1	5	1	7	10	10	9
34	10	7	8	5	2	1	2	4	8	9	7
Total	**221**	**185**	**186**	**168**	**51**	**98**	**46**	**165**	**192**	**219**	**182**

5. Discussion

To investigate more in results, we analyze to see the produced stems produced by each stemmer. Table 6 shows the unique stems produced by each stemmer. The result in this table shows interesting points. Its clear that Khoja stemmer and Lucence stemmer results are identical in most results. Both stemmers produce similar roots. It is also clear that light-based stemmers are not the same, but they all do not do challenging work in stemming. They just remove some prefixes and suffixes which success in conflating small cases and make mistakes in many other cases. For example, Assem stemmer attempt to remove the "ك" character at the end of the word "امتلك" which result in a stem that does not relate to the original word.

Table 5: Summary for stemmers conflating results.

Stemmer	Stemmed words	Unique stems	Reduction rate
Lucene	**221**	**51**	**0.909**
Khoja	**205**	**46**	**0.850**
ISRI	**221**	**98**	**0.657**
Tashphyne	**221**	**165**	**0.299**
Farasa	**221**	**168**	**0.283**
D-light	**221**	**182**	**0.208**
Assem	**221**	**185**	**0.192**
Cog	**221**	**186**	**0.187**
Light 8	**221**	**192**	**0.155**
Light 10	**221**	**219**	**0.010**

The same with Tashaphyne stemmer, which remove the prefix "ا" and suffix "ك" from the same word and result in nonsense stem "متل". It also removes both characters from the term "املاك", which result in wrong stem "ملا". ISRI stemmer has a better result in conflating terms; however, it makes several mistakes can be found in its section of the result table. For example, it removes the suffix "كم" from the word "متحكم" which result in bad stem "مت", which have no meaning with the original word. It also reduces the word "مباني", which to "مبا", which also does not relate to the original word. Cog stemmer in the other hand do minor changes to the original word in case the word is long, otherwise, it leaves the word as it is. For example, in class no 6 when the word was "عصا", the stemmer leaves it as its, but in words "عصاه" and "عصيان", it attempts to remove the suffixes "اه" and "ان", which result in wrong stem "عص". Farasa stemmer, also do slight changes to the original word by removing specific character. One mistakes of Farasa stemmer is returning the stem "طافي" for the word "لطافة". It also attempts to remove "نا" from the word "بنا" which reduce the word to the "ب" character only. D-light stemmer

also produce some wrong stems like "**ام**" in class 1, which is a result of removing the suffix "**لاك**". Many others shown in the table like "**ممل**","**احت**" and the original words are "**مملكة**", and "**احتكم**". Light 8 did not do well, but it was better than light10 in the reduction rate as light 10 did not work well in reduction as light 10 only focus on removing the definition articles, which are limited in the dataset used. Although Khoja and lucene stemmers are good, they also produce some wrong roots. for example, class no 25. One of the produced roots by both stemmers is "**ويب**", which is not related to the class words. They also failed in conflating some terms in class no 2. The words "**تستأجر**", and "**تستأجرة**", are returned as they are by both stemmers.

Table 6: Comparison between stemmers in producing the correct stem or root for the dataset

Class	Dataset	Assem	Cog	Farasa	Lucen	ISRI	Khoja	Tashaphyne	Light 8	Light 10	D-light
1	امتلك تملك ملك املاك ملاك يمتلك تمتلك ممتلكات مملكة ملوك ممالك	امتل تمل ملك امل ملا يمتل تمتل ممتلكا مملك ملو ممال	امتلك تملك ملك املاك ملاك يمتلك تمتلك ممتلك مملك ملوك ممالك	امتلك ملك أملاك مالك ممتلك مملكة	ملك لوك	ملك تمل لاك لوك	ملك لوك	متل مل ملا ممتلك مملك ممال	امتلك تملك ملك املاك ملاك يمتلك تمتلك ممتلك مملك ملوك ممالك	امتلك تملك ملك املاك ملاك يمتلك تمتلك ممتلكات مملكه ملوك ممالك	امتل تمل مل ام ملا يمتل تمتل ممتل ممل ملو ممال
2	استئجر ايجار مستأجر اجارة مستأجرين تستأجر تستأجره	استيجر ايجار مستاجر اجار استاجر	استئجر ايجار مستأجر اجار تستأجر	استئجر إيجار مستأجر اجار استأجر	أجر جور تستأجر تستأجره	ئجر يجر أجر جرة	أجر جور تستأجر تستأجره	ستئجر يجار مستأجر اجار تستأجر	استئجر ايجار مستأجر اجار تستأجر	استئجر ايجار مستاجر اجاره مستاجرين تستاجر تستاجره	استئجر ايجار مستأجر اجار تستأجر
3	تحكم حكم متحكم احتكم احتكام الاحتكام احكام	تحكم حكم متح احت احتكام احكام	تحكم حكم متحكم احتكم احتكام احكام	حكم متحكم احتكم احتكام احكام	حكم	حكم متح احت	حكم	حكم متح حت حتكام احتكام حكام	تحكم حكم متحكم احتكم احتكام الاحتكام احكام	تحكم حكم متحكم احتكم احتكام احكام	حكم متح احت احتكام احكام
4	استشار استشارة مستشار تستشير	استشار مستشار استشير	استشار مستشار تستشير	استشار استشارة مستشار تستشير	شور تستشير	شار تشر	شور تستشير	ستشار استشار مستشار تستشير	استشار مستشار تستشير	استشار استشاره مستشار تستشير	استشار مستشار تستشير
5	استئصل استئصال يستأصل استأصله	استيصل استيصال استاصل	استئصل استئصال يستأصل استأصل	استئصل استئصال استأصل	أصل	نصل أصل	أصل	ستئصل استئصال يستأصل ستأصل	استئصل استئصال يستأصل استأصل	استئصل استئصال يستاصل استاصله	استئصل استئصال يستأصل استأصل
6	عصا عصاه عصيان عصي	عصا عصيان عص	عصا عص	عصا عصيان عصي	عصي	عصا عصه عصي	عصي	عص عصا عصي	عصا عصي	عصا عصاه عصيان عصي	عصا عص
7	تائه يتوه تيه تيهان	تاء يتو تيه تيهان	تائ يتو تي	تائه يتوه تيه تيهان	توه	تئه يته تيه	توه	تائ تو يه	تائ يتو تيه	تائه يتوه تيه تيهان	تائ يتو تي
8	يغوص غوص غص غواصة	يغوص غوص غص غواص	يغوص غوص غص غواص	غاص غوص غص غواص	غوص	يغص غوص غص	غوص	غوص غص غواص	يغوص غوص	يغوص غوص	يغوص غوص غص غواص

Class	Dataset	Assem	Cog	Farasa	Lucen	ISRI	Khoja	Tashaphyne	Light 8	Light 10	D-light
									غص غواص	غص غواصه	
9	تتبع تبع يتتبع اتبع اتبعه	تتبع تبع يتتبع اتبع	تتبع تبع يتتبع اتبع	تبع تتبع اتبع	تبع	تبع	تبع	تبع بع تتبع	تتبع تبع يتتبع اتبع	تتبع تبع يتتبع اتبع اتبعه	تتبع تبع يتتبع اتبع
10	علم يعلم علما عليم معلم متعلم تعلم تعليما معلمه	علم يعلم عليم معلم متعلم تعلم تعليم	علم يعلم علما عليم معلم متعلم تعلم تعليما	علم عليم معلم متعلم تعلم تعليم	علم	علم تعلم تعل	علم علا	علم عليم معلم متعلم	علم يعلم علما عليم معلم متعلم تعلم تعليما	علم يعلم علما عليم معلم متعلم تعلم تعليما معلمه	علم يعلم عليم معلم متعلم تعلم تعل
11	ادخال دخل دخلا دخول مدخل مداخل ادخل دخيل دخلاء ادخاله ادخالها	ادخال دخل دخول مدخل مداخل ادخل دخيل دخلاء	ادخال دخل دخلا دخول مدخل مداخل ادخل دخيل دخلاء	إدخال دخل دخول مدخل مدخول أدخل دخيل	دخل	دخل دخلاء	دخل	دخال دخل دخول مدخل مداخل دخيل دخلاء	ادخال دخل دخلا دخول مدخل مداخل ادخل دخيل دخلاء	ادخال دخل دخلا دخول مدخل مداخل ادخل دخيل دخلاء ادخاله ادخالها	ادخال دخل دخول مدخل مداخل ادخل دخيل دخلاء
12	خاف يخاف خوف خافا يخافون متخوف خائف اتخاف تخيف تتخوف	خاف يخاف خوف متخوف خايف اتخاف تخيف تتخوف	خاف يخاف خوف خافا متخوف خائف اتخاف تخيف تتخوف	خافي خاف خوف متخوف خائف اتخاف أخاف تخوف	خوف خفي	خاف يخف خوف خفا تخف خنف	خوف خفي	خاف خوف متخوف خائف تخاف خيف تخوف	خاف يخاف خوف خافا متخوف خائف اتخاف تخيف تتخوف	خاف يخاف خوف خافا يخافون متخوف خائف اتخاف تخيف تتخوف	خاف يخاف خوف متخوف خائف اتخاف تخيف تتخوف
13	هرب يهرب هروب متهرب تهريب	هرب يهرب هروب متهرب تهريب	هرب يهرب هروب متهرب تهريب	هرب هروب متهرب تهريب	هرب	هرب تهرب	هرب	هرب هروب متهرب هريب	هرب يهرب هروب متهرب تهريب	هرب يهرب هروب متهرب تهريب	هرب يهرب هروب متهرب تهريب
14	تلوث لوث يتلوث تلويثها تلوثونها تلوثينها	تلوث لوث يتلوث تلويث	تلوث لوث يتلوث تلويث	تلوث لوث يتلوث تلويث تلوثينها	لوث	تلث لوث وثو	لوث	لوث وث تلوث لويث	تلوث لوث يتلوث تلويث تلوثون تلوثين	تلوث لوث يتلوث تلويثها تلوثونها تلوثينها	تلوث لوث يتلوث تلويث تلوثو
15	استيفاء اوفى وفي وفاء مستوفي ايفاء	استيفاء اوفي وف فاء مستوف ايفاء	استيفاء اوفى وف فاء مستوف ايفاء	استيفاء اوفى في وفاء مستوفي إيفاء	استيفاء وفي ايفاء	يفء وفى وفي وفء	استيفاء وفي ايفاء	استيفاء وفى في فاء مستوف يفاء	استيفاء اوفى وفي وفاء مستوف ايفاء	استيفاء اوفي في فاء مستوفي ايفاء	استيفاء اوفى في فاء مستوف ايفاء
16	عفا يعفو عفوا اعف	عفا يعفو عفو	عفا يعفو عفوا	عفا عفو اعف	عوف يعفو عفن اعفاء	عفا عفو اعف عفء يعف		عف عفو عفي اعفاء	عفا يعفو عفوا	عفا يعف عفوا اعف	عفا يعفو عفو اعف اعفاء يعف

Class	Dataset	Assem	Cog	Farasa	Lucen	ISRI	Khoja	Tashaphyne	Light 8	Light 10	D-light
	اعفي اعفه اعفينها اعفاء يعفي	اعف اعفين اعفاء يعف	اعف اعفاء يعف	اعفي اعفينها إعفاء أعفى					اعف اعفين اعفاء يعف	اعفي اعفه اعفينها اعفاء يعفي	
17	قضى يقضي قضاء افض قاض قضية قضايا	قضي يقض قضاء افض قاض قض قضاي	قضى يقض قضاء افض قاض قض قضايا	قضى قضاء افض قاضي قضية	قضي فضي	قضى يقض قضء افض قضي قضة قضا		قضى قضي قضاء فض قاض قض قضاي	قضى يقض قضاء افض قاض قضية قضايا	قضي يقضي قضاء افض قاضي قضيه قضايا	قضى يقض قضاء افض قاض قض قضا
18	استجاب يستجيب اجابة مجيب اجب استجب مجاب جواب اجوبة	استجاب استجيب اجاب مجيب اجب استجب مجاب جواب اجوب	استجاب يستجيب اجاب مجيب اجب استجب مجاب جواب اجوب	استجاب إجابة مجيب اجب استجب مجاب جواب	جوب أجب جبي	جاب تجب جبة جيب اجب سجب جوب	جوب أجب جبي	ستجاب يستجيب اجاب مجيب جب ستجب مجاب جواب اجوب	استجاب يستجيب اجاب مجيب اجب استجب مجاب جواب اجوب	استجاب يستجيب اجابه مجيب اجب استجب مجاب جواب اجوبه	استجاب يستجيب اجاب مجيب اجب استجب مجاب جواب اجوب
19	احترام محترم يحترم حرمة	احترام محترم يحترم حرم	احترام محترم يحترم حرم	احترام محترم احترم حرمة	حرم	حرم	حرم	حترام محترم حترم حرم	احترام محترم يحترم حرم	احترام محترم يحترم حرمه	احترام محترم يحترم حرم
20	احسان محسن حسن احسن استحسن استحسان	احسان محس حسن احس استحس	احس محسن حسن احسن استحسن استحس	إحسان محسن حسن أحسن استحسن استحسان	حسن	احس حسن سحس	حسن	حسا محسن حس ستحسن استحس	احس محسن حسن احسن استحسن استحس	احسان محسن حسن احسن استحسن استحسان	احس محس حسن استحس
21	جمال جميل جميله جميلات	جمال جميل جميلا	جمال جميل	جمال جميل	جمل	جمل	جمل	جمال جميل	جمال جميل	جمال جميل جميله جميلات	جمال جميل
22	لطيف لطيفه لطافه استلطفه يستلطفه يلاطفه ملاطفه	لطيف لطاف استلطف يلاطف ملاطف	لطيف لطاف استلطف يستلطف يلاطف ملاطف	لطيف طافي استلطف يستلطفه لاطف ملاطف	لطف	لطف	لطف	طيف طاف ستلطف يستلطف لاطف ملاطف	لطيف لطاف استلطف يستلطف يلاطف ملاطف	لطيف لطيفه لطافه استلطفه يستلطفه يلاطفه ملاطفه	طيف طاف استلطف يستلطف يلاطف ملاطف
23	ظرافة ظريف استظرفه	ظراف ظريف استظرف	ظراف ظريف استظرف	ظراف ظريف استظرف	ظرف	ظرف	ظرف	ظراف ظريف ستظرف	ظراف ظريف استظرف	ظرافه ظريف استظرفه	ظراف ظريف استظرف
24	استعمر مستعمر يستعمر مستعمرة مستعمرات مستعمرين استعمار يستعمرون	استعمر مستعمر مستعمرا استعمار	استعمر مستعمر يستعمر استعمار	استعمر مستعمر يستعمر استعمار يستعمرون	عمر	عمر	عمر	ستعمر مستعمر يستعمر استعمار	استعمر مستعمر يستعمر استعمار	استعمر مستعمر يستعمر مستعمره مستعمرات مستعمرين استعمار يستعمرون	استعمر مستعمر يستعمر استعمار
25	بنا	بنا يبن	بنا يبن	ب بنى	بني ويب	بنا يبن	بني ويب	بنا بني	بنا يبن ميان	بنا يبني	بنا يبن مبا بن

Class	Dataset	Assem	Cog	Farasa	Lucen	ISRI	Khoja	Tashaphyne	Light 8	Light 10	D-light
	يبني مباني بنايات بناية بنيان	مبا نايا نيان	مبان بن	مبني بناية بنيان		مبا بني		مبان نا نيان	بناي بني	مباني بنايات بنايه بنيان	
26	يكتب كتابة كتب مكاتب مكتبة كاتب	يكتب كتاب كتب مكاتب مكتب كاتب	يكتب كتاب كتب مكاتب مكتب كاتب	كتب كتابة مكتب مكتبة كاتب	كتب	كتب	كتب	كتب تاب تب مكاتب مكتب اتب	يكتب كتاب كتب مكاتب مكتب كاتب	يكتب كتابه كتب مكاتب مكتبه كاتب	يكتب كتاب كتب مكاتب مكتب كاتب
27	تركيب يركب تراكيب متراكب تركيبات متراكبة	تركيب يركب تراكيب متراكب تركيبا	تركيب يركب تراكيب متراكب	تركيب ركب متراكب متراكبة	ركب تراكيب	ركب راكب	ركب تراكيب	ركيب ركب تراكيب متراكب تركيب	تركيب يركب تراكيب متراكب	تركيب يركب تراكيب متراكب تركيبات متراكبه	تركيب يركب تراكيب متراكب
28	نظر ينظر منظار نظارة يناظره متناظر نظير نظائر	نظر ينظر منظار نظار يناظر متناظر نظير نظاير	نظر ينظر منظار نظار يناظر متناظر نظير نظائر	نظر منظار نظار ناظر متناظر نظير	نظر	نظر ناظر	نظر	ظر نظر منظار نظار ناظر متناظر ظير ظائر	نظر ينظر منظار نظار يناظر متناظر نظير نظائر	نظر ينظر منظار نظاره يناظره متناظر نظير نظائر	نظر ينظر منظار نظار يناظر متناظر نظير نظائر
29	حماقه حمقى احمق	حماق حمقي احمق	حماق حمقى احمق	حماق حمقى احمق	حمق	حمق حمقى	حمق	حماق حمقى حمق	حماق حمقى احمق	حماقه حمقي احمق	حماق حمقى احمق
30	سهوله سهل يستسهل مستسهل استسهال	سهول سهل استسهل مستسهل استسهال	سهول سهل يستسهل مستسهل استسهال	سهل استسهل مستسهل استسهال	سهل	سهل	سهل	سهول سهل يستسهل مستسهل استسهال	سهول سهل يستسهل مستسهل استسهال	سهوله سهل يستسهل مستسهل استسهال	سهول سهل يستسهل مستسهل استسهال
31	اعترف يعترف اعتراف معترف اعترفي	اعترف يعترف اعتراف معترف	اعترف يعترف اعتراف معترف	اعترف اعتراف معترف	عرف	عرف	عرف	عترف عتراف معترف	اعترف يعترف اعتراف معترف	اعترف يعترف اعتراف معترف اعترفي	اعترف يعترف اعتراف معترف
32	اوصى يوصي وصية وصاية اوصاه اوصاها وصاها وصاه	اوصي يوص وص وصا اوص	اوصى يوص صي صا اوصا	أوصى وصية وصاية اوصا وصا	وصي	وصى يوص وصة وصي وصه اوص وصا	وصي	وصى وصي صي صا وصا وص	اوصى يوص وصية وصا اوصا	اوصي يوصي صيه صايه اوصاه اوصاها صاها صاه	اوصى يوص صي صا او اوص
33	توقف استوقف متوقف مستوقف واقف وقوف مواقف قف قفي قفا	توقف استوقف متوقف مستوقف واقف قوف مواقف قف قفا	توقف استوقف متوقف مستوقف اقف قوف مواقف قف قفا	توقف استوقف متوقف مستوقف واقف وقوف موقف قف قفي قفا	وقف	وقف توقف قف قفي قفا	وقف	وقف ستوقف متوقف مستوقف قف قوف مواقف	توقف استوقف متوقف مستوقف واقف وقوف مواقف قف قفي قفا	توقف استوقف متوقف مستوقف اقف قوف مواقف قف قفي قفا	توقف استوقف متوقف مستوقف اقف قوف مواقف قف قفا

Class	Dataset	Assem	Cog	Farasa	Lucen	ISRI	Khoja	Tashaphyne	Light 8	Light 10	D-light
34	تقرير تقارير قرر قرار يقرر قرار اقرار يقررون تقرر قررت	تقرير تقارير قرر قرار يقرر اقرار تقرر	تقرير تقارير قرر قرار يقرر اقرار تقرر قررت	تقرير قرر قرار إقرار تقرر	قرر قور	قرر	قرر قور	قرير تقارير قرر قرار	تقرير تقارير قرر قرار يقرر اقرار تقرر قررت	تقرير تقارير قرر قرار يقرر اقرار يقررون تقرر قررت	تقرير تقارير قرر قرار يقرر اقرار تقرر

6. Conclusion

In conclusion, this paper highlights the importance of preprocessing and stemming for Arabic IR. The study provides a comprehensive analysis of the existing literature on Arabic preprocessing and stemming techniques, identifying their limitations and challenges. The results of the study indicate that root-based stemmers work well in conflating similar terms and reducing the required space for indexing, however they still have some limitations in stemming some terms, but they are better than light-based stemming techniques. The light-based stemmers in the other hand do small effort in conflating similar terms and most of the time they return the original word or do small modification to it. The findings of this study underscore the need for continued research in this area to overcome the challenges posed by the language's complex morphology to enhance the Arabic IR. Future work will include evaluating the effect of stemmers in the retrieval results.

7. References

[1] C. D. Manning, *An introduction to information retrieval*. Cambridge university press, 2009.

[2] J. Atwan, M. Mohd, H. Rashaideh, and G. Kanaan, "Semantically enhanced pseudo relevance feedback for Arabic information retrieval," *J Inf Sci*, vol. 42, no. 2, pp. 246–260, 2016.

[3] S. R. El-Beltagy and A. Rafea, "An accuracy-enhanced light stemmer for arabic text," *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 7, no. 2, pp. 1–22, 2010.

[4] H. Abu-Salem, M. Al-Omari, and M. W. Evens, "Stemming methodologies over individual query words for an Arabic information retrieval system," *Journal of the American Society for Information Science*, vol. 50, no. 6, pp. 524–529, 1999.

[5] A. Alnaied, M. Elbendak, and A. Bulbul, "An intelligent use of stemmer and morphology analysis for Arabic information retrieval," *Egyptian Informatics Journal*, vol. 21, no. 4, pp. 209–217, 2020.

[6] L. S. Larkey, L. Ballesteros, and M. E. Connell, "Improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis," in *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, 2002, pp. 275–282.

[7] A. Y. Muaad *et al.*, "Arabic document classification: performance investigation of preprocessing and representation techniques," *Math Probl Eng*, vol. 2022, pp. 1–16, 2022.

[8] K. Darwish, W. Magdy, and others, "Arabic information retrieval," *Foundations and Trends® in Information Retrieval*, vol. 7, no. 4, pp. 239–342, 2014.

[9] H. M. Al-Barhamtoshy, K. M. Jambi, S. M. Abdou, and M. A. Rashwan, "Arabic documents information retrieval for printed, handwritten, and calligraphy image," *IEEE Access*, vol. 9, pp. 51242–51257, 2021.

[10] A. A. Rafea and K. F. Shaalan, "Lexical analysis of inflected Arabic words using exhaustive search of an augmented transition network," *Softw Pract Exp*, vol. 23, no. 6, pp. 567–588, 1993.

[11] M. R. Al-Maimani, A. Al Naamany, and A. Z. A. Bakar, "Arabic information retrieval: techniques, tools and challenges," in *2011 IEEE GCC Conference and Exhibition (GCC)*, 2011, pp. 541–544.

[12] H. Froud, A. Lachkar, and S. A. Ouatik, "Arabic text summarization based on latent semantic analysis to enhance arabic documents clustering," *arXiv preprint arXiv:1302.1612*, 2013.

[13] H. A. Taher, M. H. Abdulameer, and B. Mahdi, "INFORMATION RETRIEVAL SCHEME VIA SIMILARITY TECHNIQUE," *International Journal on "Technical and Physical Problems of Engineering" (IJTPE)*, vol. 14, no. 51, pp. 375–379, 2022.

[14] B. Al-Shargabi, F. Olayah, and W. A. L. Romimah, "An experimental study for the effect of stop words elimination for arabic text classification algorithms," *International Journal of Information Technology and Web Engineering (IJITWE)*, vol. 6, no. 2, pp. 68–75, 2011.

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

[16] J. Atwan, M. Mohd, and G. Kanaan, "Enhanced arabic information retrieval: Light stemming and stop words," in *Soft Computing Applications and Intelligent Systems: Second International Multi-Conference on Artificial Intelligence Technology, M-CAIT 2013, Shah Alam, August 28-29, 2013. Proceedings*, 2013, pp. 219–228.

[17] I. A. El-Khair, "Effects of stop words elimination for Arabic information retrieval a comparative study, study," *International Journal of Computing and Information Sciences, Decembers*, 2006.

[18] K. Taghva, R. Elkhoury, and J. Coombs, "Arabic stemming without a root dictionary," in *International Conference on Information Technology: Coding and Computing (ITCC'05)-Volume II*, 2005, pp. 152–157.

[19] S. Ben Guirat, I. Bounhas, and Y. Slimani, "Enhancing hybrid indexing for Arabic information retrieval," in *Computer and Information Sciences: 32nd International Symposium, ISCIS 2018, Held at the 24th IFIP World Computer Congress, WCC 2018, Poznan, Poland, September 20-21, 2018, Proceedings 32*, 2018, pp. 247–254.

[20] A. El Mahdaouy, S. O. El Alaoui, and E. Gaussier, "Improving Arabic information retrieval using word embedding similarities," *Int J Speech Technol*, vol. 21, pp. 121–136, 2018.

[21] L. S. Larkey, L. Ballesteros, and M. E. Connell, "Light stemming for Arabic information retrieval," in *Arabic computational morphology: knowledge-based and empirical methods*, Springer, 2007, pp. 221–243.

[22] Shereen Khoja, "Khoja Stemmer." Free Software Foundation, 2002. Accessed: Aug. 19, 2023. [Online]. Available: http://zeus.cs.pacificu.edu/shereen/ArabicStemmerCode.zip

[23] K. Darwish, "Building a shallow Arabic morphological analyser in one day," in *Proceedings of the ACL-02 workshop on Computational approaches to semitic languages*, 2002.

[24] S. Ghwanmeh, G. Kanaan, R. Al-Shalabi, and S. Rabab'ah, "Enhanced algorithm for extracting the root of Arabic words," in *2009 sixth international conference on computer graphics, imaging and visualization*, 2009, pp. 388–391.

[25] S. Ben Guirat, I. Bounhas, and Y. Slimani, "A hybrid model for Arabic document indexing," in *2016 17th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, 2016, pp. 109–114.

[26] M. G. Syarief, O. T. Kurahman, A. F. Huda, and W. Darmalaksana, "Improving Arabic stemmer: ISRI stemmer," in *2019 IEEE 5th International Conference on Wireless and Telematics (ICWT)*, 2019, pp. 1–4.

[27] H. Alshalabi, S. Tiun, N. Omar, E. abdulwahab Anaam, and Y. Saif, "BPR algorithm: New broken plural rules for an Arabic stemmer," *Egyptian Informatics Journal*, vol. 23, no. 3, pp. 363–371, 2022.

[28] H. Khafajeh, N. Yousef, and G. Kanaan, "Automatic query expansion for Arabic text retrieval based on association and similarity thesaurus," in *Proceedings he European, Mediterranean & Middle Eastern Conference on Information Systems (EMCIS), Abu Dhabi, UAE*, 2010.

[29] A. F. Smeaton and C. J. van Rijsbergen, "The retrieval effects of query expansion on a feedback document retrieval system," *Comput J*, vol. 26, no. 3, pp. 239–246, 1983.

[30] Y. Kadri and J.-Y. Nie, "Effective stemming for Arabic information retrieval," in *Proceedings of the International Conference on the Challenge of Arabic for NLP/MT*, 2006, pp. 68–75.

[31] Y. Jaafar, D. Namly, K. Bouzoubaa, and A. Yousfi, "Enhancing Arabic stemming process using resources and benchmarking tools," *Journal of King Saud University-Computer and Information Sciences*, vol. 29, no. 2, pp. 164–170, 2017.

[32] M. Mustafa, A. S. Aldeen, M. E. Zidan, R. E. Ahmed, Y. Eltigani, and others, "Developing two different novel techniques for Arabic text stemming," *Intell Inf Manag*, vol. 11, no. 01, p. 1, 2019.

[33] H. Alshalabi, S. Tiun, N. Omar, F. N. AL-Aswadi, and K. A. Alezabi, "Arabic light-based stemmer using new rules," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 9, pp. 6635–6642, 2022.

[34] I. A. Al-Kharashi and M. W. Evens, "Comparing words, stems, and roots as index terms in an Arabic information retrieval system," *Journal of the American Society for Information Science*, vol. 45, no. 8, pp. 548–560, 1994.

[35] M. M. Syiam, Z. T. Fayed, and M. B. Habib, "An intelligent system for Arabic text categorization," *International Journal of Intelligent Computing and Information Sciences*, vol. 6, no. 1, pp. 1–19, 2006.

[36] Y. A. Al-Lahham, "Index Term Selection Heuristics for Arabic Text Retrieval," *Arab J Sci Eng*, vol. 46, no. 4, pp. 3345–3355, 2021.

[37] S. Gadri and E. Neuhold, "Developing a Multilingual Stemmer for the Requirement of Text Categorization and Information Retrieval," *International Journal on Electrical Engineering and Informatics*, vol. 14, no. 2, pp. 291–310, 2022.

[38] A. A. R. Mohamed, C. Ouni, S. M. Eljack, and F. Alfayez, "Information Retrieval Systems: Between Morphological Analyzers and Systemming Algorithms," *IJCSNS*, vol. 22, no. 3, p. 375, 2022.

[39] samer yaseen, "Arabic Stemming to Select Index Terms.," *Dataset to test Arabic stemmers*, vol. 1, no. 1. Mendeley Data, Jul. 11, 2023. doi: 10.17632/w4y9pb9w8n.1