Modern Applied Science; Vol. 13, No. 6; 2019
ISSN 1913-1844 E-ISSN 1913-1852
Published by Canadian Center of Science and Education

Retrieving Arabic Textual Documents Based on Queries Written in Bahraini Slang Language

Ayat Amin Al-Jarrah[1], Ghassan Kanaan[1] & Mustafa Abdel-Kareem Ababneh[1]

[1] Computer Science and Informatics, Amman Arab University, Amman, Jordan

Correspondence: Ayat Amin Al-Jarrah, Amman Arab University, Computer Science and Informatics, Amman, Jordan. E-mail: toootoo89@yahoo.com

Received: April 1, 2019 Accepted: April 28, 2019 Online Published: May 23, 2019

doi:10.5539/mas.v13n6p44 URL: https://doi.org/10.5539/mas.v13n6p44

Abstract

Nowadays, the most used language is the colloquial language not the classical language. It is widely used in many nations. The kingdom of Bahrain had the largest share in the spread of the colloquial language, which becomes the trader's language and the language of the social communication too. It became so popular that its usage starts dominating the daily conversations. In this research, we will create algorithm to enhance the process of information retrieval in Arabic slang language of the Gulf. In this algorithm, we put some special Bahraini rules to convert queries from Slang Bahraini to a classical language. In addition, we will apply this algorithm on the Bahraini colloquial language. After making an evaluation for the system relying on the results of three main aspects recall, precision, and F-measure, we noticed that the results of precision about 0.64 for both researches slang and classical, which gives a great indication that the system supports searching in Bahraini slang language. The purpose of this research is to improve the Information Retrieval system field. In addition, it will save the time and the effort of the researchers of the Bahraini colloquial language.

Keywords: bahraini arabic slang language grammar, slang Language, classical language

1. Introduction

1.1 Introduce the Problem

Nowadays, the use of Slang has become very common in several fields; social, academic, scientific, practical, and political fields. Daily conversations (Almeman & Lee, 2013), SMS, e-mails, and social media (Harrag and et al., 2011; Duwairi and et al., 2007; Al-Shalabi and et al., 2003) are devoid of standard Arabic and are bounded to Slang.

Arabic is not new language since it has precedes Islam (Azmi, 2017). Unlike any language, Arabic is a Semitic language. It is the fourth most wildly used language in the world, with a usage percentage of 5% of people around the world (Kadri & Nie, 2006).

As long as the Arabs new generations are missing the knowledge of many standard terms of Arabic, they will suffer from obscurity of these terms as well as constant decreasing usage of the same terms. Some studies have indicated the absence of standard Arabic speakers from native speakers of Modern Standard Arabic (MSA) (Mutahhar & Watson, 2002).

As observed, there are 12 Arabic dialects used in 28 countries today. The most popular one of these dialects is the Egyptian dialect. The diversity of these dialects is increasing in many countries and it is used instead of the Standard Arabic (Rogerson, 2008).

Information Retrieval (IR) science includes many studies which help users in retrieving information, that only concerns them (Ghoneim, 2008; Sanan and et al., 2008). Unlike English language dialects studies, the researches for Arabic ones are kind of limited and miniscule (Alamlahi & Ahmed, 2011).

Well known that there are many characteristics, that distinguish the Arabic language from other languages such as the direction of writing from right to left (Benjiba and et al., 2007; Nwersi, 2008), and the shape of the character which changes according to its position in the sentence (Ghoneim, 2008; Sanan and et al., 2008; Al-Shalabi and et al., 2009). Many algorithms used to extract the root of the word, then search within the dictionary about its meaning (Goweder and et al., 2008; Al-Shalabi and et al., 2005; Ahmed and et al., 2008), but so far no algorithm has been able to find the full and the complete stem of the Arabic words (Goweder and et al., 2008).

1.2 Explore Importance of the Problem

Due to the importance of Arabic language and its wide spread in many countries, it has spread in the Islamic and non-Arabic-speaking ones. Arabic language is at the second rank with about 60 countries speaking, it includes 22 countries of the international organization of Arabic-speaking language, and other countries are between Asia and Africa. On the other hand, English language is classified as the first rank in the world with about 100 countries speaking it. Moreover, Arabic language is at the fourth rank according to the number of countries that speak it as their mother language (Albalooshi and et al., 2011).

2. Previous Studies

It is very significant to provide relevant information for users since this will save both time and effort for all sides. Therefore, any Improvement in the models of IR in representing them through computers as numerical ones will be an improved step in IR modernity. Some paper shows a new method of IR that may appear as a vector of multi distances between concepts of queries and distances of the same documents concepts. As a result, it differentiates the relevant results between queries and documents by making a match of distances between each two concepts, in the query with the distances among the similar concepts in every corpus documents. Moreover, this research is using an example to reveal the ability of the new model in identifying the two or more documents depending on the given query, whichcannot be done by using the vector space model as the documents consist of the same frequent term and the same number of total concepts (Wijewickrema&Ratnayake, 2013).

It is noticeable that Retrieving Information is a complicated process because it is a wasting of time and effort precisely in case that it is used in wide data through internet or in large data. As a result, a lot of algorithms and steps were made to make it an easy one. Nevertheless, retrieving is in continuity towards being very complicated. Applications are used in order to add information, delete them, or even change them, for instance the Book Library System, and the Commercial Document Retrieval Services. This constrains the different types of both data and algorithms that are really built to be used in the IR. In some research, the strategy of VSM (Vector Space Model) is a good method for IR. To begin with, we try to discover the proportions of similarities in scores usage for each item. Then, we compare the results with each other. After that, the cosine measure is applied to compute the similarity measure, and in order to check the angle of both the documents vector and the vectors query whereas VSMs are mainly built on geometry. Every term must have its own dimension in a multiple space of dimensions, documents and queries. Therefore, we found out that it is easier to retrieve the information based on the similarity to produce a very professional and efficient strategy for information retrieval (Ogheneovo & Japheth, 2016).

There isSome Method on Arabic Dialects such as; the Cross Lingual Arabic Blog Alerts (COLABA), which is, defined as very inspirational work that aims to find out many techniques and methods to help the Arabic dialects. It was institute to have data from Arabic aspects such as; blogs, chat rooms, social media, and forums. In addition, it should be recognized that the Arabic dialects are clearly used in the previous forums. COLABA project gave attention to the IR as a very used way to indicate Arabic slang. The IR has a very big role in recovering the dialectical information at the same time with the MSA standard format. Therefore, it allows users to recover contents as far as you can tell. To change the concepts of query from MSA to the assigned dialect, a term from MSA was built to find the words harmony (Diab and et al., 2010).

Moreover, researchersproduce an algorithm to turn the Yemeni language "Sana'a Dialects" into a modern standard language. This algorithm was based on a strategy of removing vowels, suffixes, and prefixes. After that, they extract the stem and comparing it with the colloquial stems used in the Yemini language and the stems of MSA. Finally, finding the texts in the MSA language. After practicing the algorithm on many of the Yemeni texts, they found that 16.29% of these words contain distorted suffix, 2.16% contain distorted stem and 0.70% contain distorted suffix. (Al_Gaphari & Al_Yadoumi, 2006).

Another research is done on gulf dialect because this dialect is a common dialect in many countries; (Kuwait, Bahrain, Qatar, UAE, some eastern regions of Saudi Arabia, and some areas of southern Iraq). Many of the Gulf words have become obsolete; some of them due to their non-Arab origins. This algorithm adopted Stem Extraction from the slang Gulf word with the assumption that the root of the word is three letters and by removing the affixes of the word. Then by comparing it with non-Arabic words and stop words. After the implementation of the above-mentioned algorithm on the Arabic Gulf dialects, the accuracy of the new stem was in a better state than its state on other roots. In addition, the affixes were very few and resulted from some modifications to the characters by the Gulf people (Abuata & Al-Omari, 2015).

Some approach design in IR system to retrieve Saudi Arabia dialect to modify an MSA stemming and a set of colloquial to MSA conversion rules that are lexicon based and the result of this method is given after testing 44 queries over 1400 documents where 84.3% precision, and 96.5% for recall. (Azmi & Aljafari, 2015).

Activities in the social media and network became very noticeable and commonly the slang one is used. That is why many researches were done and one of them mainly uses Slang Sentimental Words,where Idioms Lexicon (SSWIL) are constructed. A Gaussian Kernel Support Vector Machine (SVM) classifier is used for Arabic slang to classify comments of social media. Several Facebook comments are used too to test the performance of the suggested classifier; the accuracy rate reached about 86.86% (Elmasry and et al., 2014).

This research proposed a SVM-based classifier for Arabic slang, and it applied sentiment analysis to classify young people's comments on Facebook. The process of classifying consisted of three main phases: The preparation of Arabic data comments, Data preliminary processing and Data grouping.

Soliman & Hedar (2014) research focuses on the comments of social media's users. Young Arabs on social media; on Facebook post and comments, and Twitter's Twits and reviews use (SSWIL). Furthermore, it improves the classification task to obtain its highest accuracy rates of 86.86% instead of the previous rate of 75.35%. Therefore, the records improved to become 88.63% for precision and 78 for recall instead of (82.4% – 59.33 %) respectively (Soliman & Hedar, 2014).

Another study is a sociolinguistic investigation of Saudi Youths Slang. Many countries have several researches and projects on the youths slang except Saudi Arabia, which is destitute of such researches on such a linguistic phenomenon. This study analyze the social factors, puts a great focus on the sources from which the Saudi youths acquire their new slang expressions, and inspects the reasons behind using such expressions. Moreover, the study discusses the major focus of Saudi Arabia youths, and what attracts them to use such slang. Finally, the study concludes that Age and Saudi Youths slang are interactive. Which means that the younger the person is, the higher the level of hi/her association with slang. Moreover, it indicates that men use more slang expressions than women do. In addition, it shows the diversion of the slang topics that attracts the attention of Saudi men and women; particularly, sexual topics. Furthermore, the study provides several typical examples of current Saudi youth slang (Gomaa and et al., 2015).

3. Methodology

The basic purpose for this research is to design a useful system that works to make restoring for Arabic languages information by using mainly the Bahraini slang. In addition, it works by normalizing the entered query by applying some rules, searching, and restoring a group of documents. These documents were collected from the social media websites like Twitter, Facebook, Instagram, Whats app, etc.

3.1 System Framework

Figure (1) represents the framework of the Bahraini slang Arabic Retrieval system. In addition, it clarifies the significant steps that the system applies to restore the information by using the Bahraini slang. Therefore, the goal of this framework is to normalize the query that is written in the Bahraini slang. In addition, to achieve this goal, the system should contain two sections to deal with the written query after checking the grammatical structure of the sentence by enforcing the special grammar for the system. Thus, if the written query is correct grammatically, it moves to the next step. However, if the query is refused, the user will be asked to insert a new one.

The first approach stands by keeping the query in its slang form, and it deletes the stop words. Then, it searches for the query in the written slang documents. However, the other approach works by turning the slang query into a classical form by applying the rules and by searching for the unknown words in the non-Arabic list and the look up table. It deletes the stop words, and it searches for it in the written document of the classical form. Finally, the resulted documents for both kinds are restored and appeared to the user.

3.2 Bahraini Arabic Slang Grammar

The system of restoring the information by using the slang Bahraini depends totally on the grammar of the Arabic Slang Grammar (ASG). It checks the query that the user put, if it agrees with the grammar, it will be processed, and the rest of the process will be done. However, if it did not match the rules, and the system refused it, it will ask the user to enter a new query.

In this stage, the type of any word in the query is determined whether it is noun, verb, question mark, preposition,etc.

3.3 Definition of the Grammar

In this research, we will use special grammar in the Bahraini slang language. These rules consist of three parts:

- Noun clause (Ns): the most important item in it is the noun (N).

- The verbal clause: the most important item in it is the verb (V).

- Interrogative sentence: the most important item in it is the question mark (Q).

* The content of the grammars of the slang Bahraini language are shown in Figure (2), and the list of the used grammars are shown in Table (1).

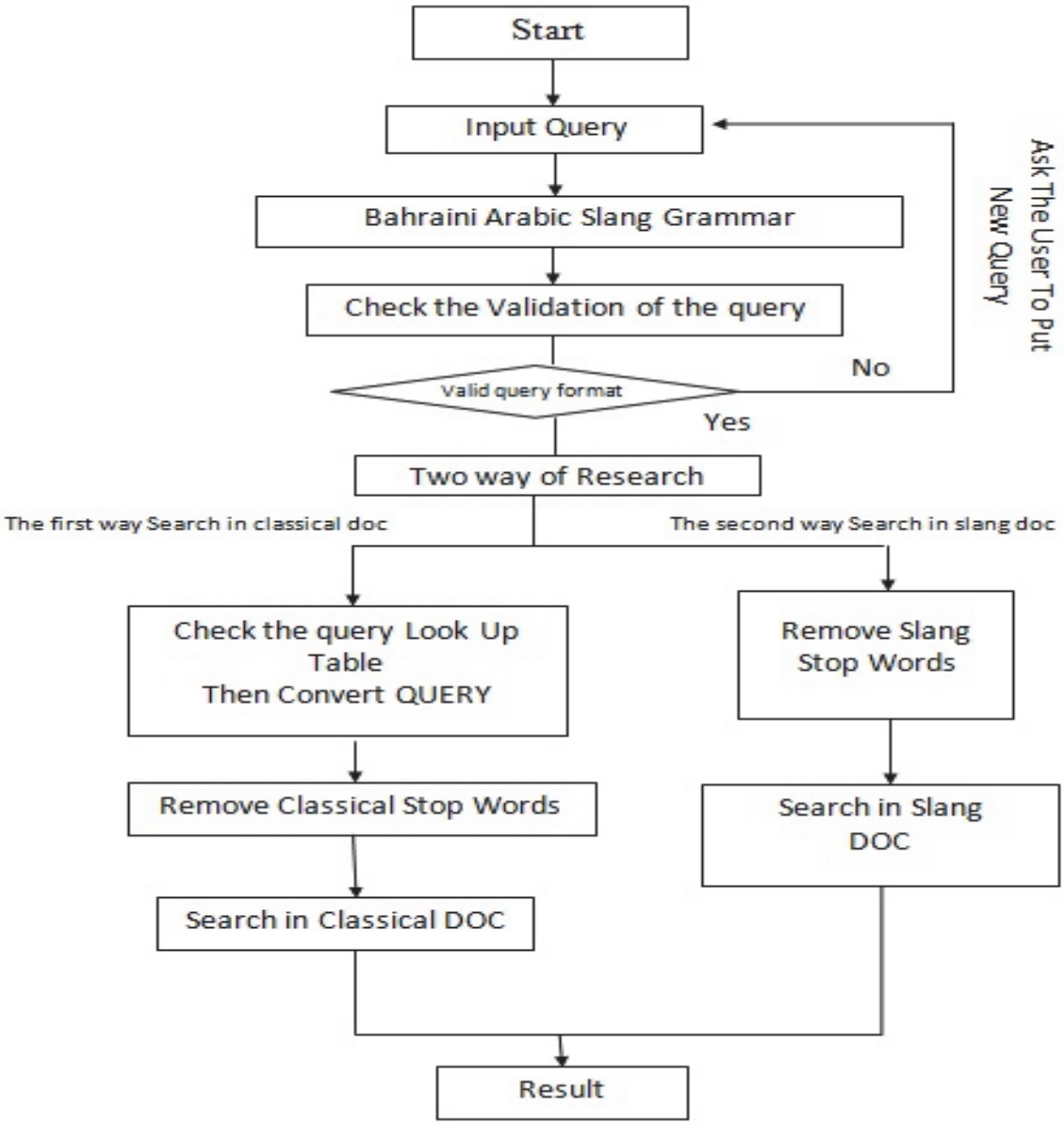

Figure 1. The Framework

Table 1. The Abbreviations of the Grammar

Abbreviation	Meaning
Pre1	List of prefix for present verb {ت | ن | ي | أ}
Pre2	List of prefix for imperative verb {ا}
Suf 1	List of suffix for present verbs {ون | ين}
Suf2	List of suffix for past verbs {و | نا | تي | ت}
Suf3	List of suffix for Imperative verbs {ي | وا | و}
PreLN	List of prefix letter of nouns{ال | لل | وال | بال | وللل}
Suf LN	List of suffix letter of nouns

Number table (No)	{ات | ان | ين | ون | ك | ج | كم | وكم | ونه | تيه | تني} واحد | اثنين | ثلاثة | اربعة | خمسة | ستة | سبعة | ثمنية | تسعة} |عشرة | احدعش | اثنعش | ثلثطعش | اربعطعش | خمسطعش | ستطعش | سبعطش | منتعش | تسعطش | عشرين}
Jar table (Jar)	{من | الى | عن | على | مع | ب | في | فال | ل | عال}
Calling table (Cal)	{ يا }
Zarf table (Zarf)	{ عند | داخل | فالنص | ورا | ينب | قبل | بعد | تحت | بالوسط | في وسطهم | جدام | فوق | شهر | يوم | دقيقة | لحظة | ثانية}
Mawsool names table (MS)	{ اللي | الي }
Jazem table (JZ)	{ما | لا }
Naseb table (NS)	{اذا | مو | علشان }
Question (Q)	Question terms{ | من | منو | شلون | وين | اي واحد | ليش | شنو | هل | شحقه | جم | بجم | بأي | متى}
SufQ	Suffix of question {كم| ن| ك| ج}
SN	{ذي | هاذي | هاي | مناك| هناك| هاذيله | هذلين }
Kan	{كان | صار | ظل | للحين | بقى | لا | ليحين| مدام | جان صار }
Pro	{انت | انتي | انا | اهمه | اهما | انتوا | انتو | انتم | اهو | اهي | اهم | احنا }
Fn	{ابو | اخو | حماي }
En	{ان | كأنه | جنه | لكن | ياريت | جان زين }
Negative (Neg)	{ما| مش| مو| لا}
Atf	{او | و}

Table (2) shows to us the meaning of symbols that were used in the Bahraini Slang Language Grammar Figure (2).

Table (2). The meaning of symbols shows in Figure (2)

Symbols	Meaning
APP	The group of symbols for the grammar shows in Table (1)
PreTN	The tools preceding nouns
PreTV	The tools preceding verbs
V1	Past verb with affixes
V2	Present verb with affixes
V3	Imperative verbs with affixes
N1	Noun with affixes
Q1	Question tool with affixes
Nss	All the noun's structure

APP= {Pre1, Pre2, Suf1, Suf2, Suf3, PreLN, SufLN, No, Jar, Cal, Zarf, MS, JZ, NS, Q, SufQ, SN, EN, Kan, Pro, FN, V1, V2, V3, N, Neg, Atf }

PreTN = { No, Jar, Cal, Zarf }

PreTV = { MS, JZ, NS }

V1 ⟶ V | V Suf2

V2 ⟶ V| Pre1 V| Pre1 V Suf 1

V3 ⟶ V| Pre2 V| Pre2 V Suf 3

N1 ⟶ N |PreLN N| PreLN N SufLN| N SufLN

Q1 ⟶ Q1 | Q Suf Q

NSS ⟶ N1 | PreTN N1

VSS ⟶ V1 |V2 |V3 | PreTV V1| PreTV V2| PreTV V3

NC ⟶ Pro N1 V1 NSS | Pro N1 V2 NSS | Pro N1 Neg V1 NSS | Pro N1 Neg V2 NSS | Pro N1 NSS | Pro N1 Atf N1 | N1 V1 NSS | N1 V2 NSS | N1 V3 NSS | N1 Neg V1 NSS | N1 Neg V2 NSS | N1 V1 Atf V1 | N1 V2 Atf V2 | N1 V3 Atf V3 | N1 Neg V1 Atf Neg V1 | N1 Neg V2 Atf Neg V2 | N1 VSS V2 |N1 VSS NSS |SN N1 NSS | SN N1 Atf N1 | SN N V1 NSS | SN V1 Atf V1| SN V2 Atf V2| SN V1 NSS | SN V2 NSS | SN Neg V1 NSS | SN Neg V2 NSS|EN N1 NSS |EN SN NSS |EN V2 |N1 NSS |EN N1 V2 | EN N1 Neg V2 |Kan N1 N1 NSS| Kan N1 V2 NSS| Kan V2 N1 NSS|FN N1 |SN FN |FN V2 NSS |Pro V1 |Pro V2 |Pro V3 | Pro Neg V1 |Pro Neg V2 |Pro V1 Atf V1 | Pro V2 Atf V2 | Pro V3 Atf V3 | Pro Neg V1 Atf Neg V1 | Pro Neg V2 Atf Neg V2 | Pro V1 NSS | Pro V2 NSS | Pro V3 NSS | Pro Neg V1 NSS | Pro Neg V2 NSS | Pro MS V1 NSS | Pro MS V2 NSS | Pro MS Neg V1 NSS | Pro MS Neg V2 NSS

VC ⟶ V1 N1 NSS | V2 N1 NSS | Neg V1 N1 NSS | Neg V2 N1 NSS | V1 NSS N1| V2 NSS N1 | Neg V1 NSS N1| Neg V2 NSS N1 |VSS NSS|V1 FN N1 VSS | V2 FN N1 VSS |V3 N VSS |VSS V2

QC ⟶ Q1 VSS | Q1 VSS N1| Q1 VSS SN|Q1 N1 V2 | Q1 V2 N | Q1 N1 N1 NSS | Q1 Pro N1 N1| Q1 V1 N1 NSS | Q1 V2 N1 NSS | Q1 V2 V1 NSS | Q1 V2 V2 NSS |Q1 NSS |Q1 N1 VSS N1 |Q1 N1 N1 V2 | Q1 N1 N1 V1 N1 | Q1 N1 N1 V2 N1

Figure (2). Bahraini Slang Language Grammar

3.4 Term Type Determination

To check the structural structure of the written query, and match it with the grammars, we must have a group from the clear structures of the query in order to compare the written queries with this one, whereas the words that build a sentence, will be either a noun, a question mark, a verb, or a group of the stop words. Figure (2) shows this rule.

The verb is divided into many types and it has a group of affix that was clarified the same as the question marks and the nouns too which have a group of affixes that were shown in Table (1) too briefly.

3.4.1 Verb Type Determination

The verb is one of the sentence components where it is either past, present, or imperative. It has a group of various affixes according to the verb's kind past, either present, or imperative. In addition, the number of the affixes letters either one letter or two or none. In addition, this pseudo code is for distinguishing the verb's type.

```
If (prefix of term in the list of Pre1 || suffix of term in the list of Suf2)
Then
Term_type = present verb
Else if (prefix of term in the list of Pre2 || suffix of term in the list of Suf3)
Then
Term_type = imperative_verb
Else if (suffix of term in the list of Suf2)
```

Figure (3). Pseudo code of Arabic verbs

The present verb is connected with suffix and prefix like (**تلبس**) where it is preceded by a prefix, but it is not followed by a suffix. However, (**تلبسين**) is preceded by a prefix and a suffix. The same as the imperative verb that may be followed by prefix or a suffix like (**اشرب**) where it is preceded by a prefix, but the example of (**شرﻮا**) it is followed by a suffix.

When it comes to the past verb, it is different from both the present and the imperative. The past does not relate with any prefix, but it relates with only a suffix like (**كتبت**).

If the verb has more than four letters it may have many infixes. hese are some cases in the past tense (**اشتغلت**), the present verb (**يشتغل**), and the imperative verb (**اشتغل**). The verb also has some equipments that are defined in Table (1) that contains relative pronouns, the conditional pronouns, and the particles.

It is clear that verbs may be preceded by either relative pronouns (MS) or sealing tools (JZ) and sometimes by monument tools (NS). However, this part was considered as main one of the grammar rules section.

3.4.2 Noun Type Determination

Nouns are all the known words that include the proper nouns like the names of people, animals, places and cities…etc. In general, most of the nouns are preceded by (**ال,ﻟﺎ, والﻟﺎ**). Also, the noun is preceded sometimes by numbers, prepositions, adverbs of time, adverbs of place and calling tools.

```
If (prefix of term in the list of PreLN|| suffix of term in the list of SufLN)
Then
Term_type = noun
Else if (term is preceded by No, Jar, Cal or Zarf)
Term_type= noun
```

Figure (4). Pseudo code of Arabic nouns

3.4.3 Look Up Table

It is a list that contains many Bahraini slang words, which are not existed in the classical language. In this list, we have all the synonyms of each classical word. Thus, we use it to get some help in the process of mapping.

3.5 Validation

This section is one of the information restoring system of the slang language that aims to check each input query by checking the grammar that were previously entered where the structure for both the input query is compared to the existed structure inside the grammar. If its arrangement is right, the system moves to the next step. But if it was wrong, it gives the user an order to write a new query. Table (3) shows to us the used signs to describe each type of the Arabic words.

Table 3. Symbols used to describe each type of Arabic words

Words Type	Symbol
Past Verb (V1)	*
Present Verb (V2)	*?
Imperative Verb (V3)	*!
Noun (N)	N
Question (Q)	$
Numbers, Jar, Calling, Zarf (PreTN)	&
Mawsool, Jazem, Naseb (PreTV)	@
Sign Name (SN)	#
EN w Akhwatha (EN)	?
KAN w Akhwatha (KAN)	P
Pronoun (Pro)	!
Five Name (FN)	
Negative (Neg)	%
Atf	*%

The explain of the validation:

Table 4. The Format of Arabic Statement

Example 1	words	شحقه	اليهال	يصيحون
	symbol	$	N	*?
Example 2	words	يصيحون	اليهال	شحقه
	symbol	*?	N	$

If the user input "**شحقه يهال يصيحون**", After making a comparison between this structure and the known formula in the grammar (ASG), the system look to see if this format is right or not. Thus, the system accepted this query, but after checking another query "**يصيحون يهال شحقه**", this structure does not match any structure. Therefore, it will be refused and the user will be asked again to write a query in another way.

3.6 Mapping Slang Language to Classical Language

This step is considered as one of the Slang Arabic Information Restoring System that mainly aims to convert words from slang into classical by making the process of mapping for all the query words. By making mapping for verbs, the affixes that are related to the slang verbs are converted to the equivalent affixes that are related to the classical form. In addition, the look up table are used to convert some Arabic Bahraini Slang words. Although this process is expensive, but it gives a lot of beneficial results.

3.6.1 Verbs Mapping

To convert verbs from slang to classical, the prefixes of slang are exchanged with the prefix of the classical and the suffix of the slang is exchanged with the suffix of the classical.

Examples of Affixes in the Past, Present and Imperative Tense

Table (5, 6, 7 and 8) Examples of Affixes in the Past, Present and Imperative Tense to explanation how mapping Slang Language to Classical Language or vice versa.

Table 5. Examples of Slang Prefix for Present verb and their equivalent classical ones

Slang	Present Prefix	Classical	Present Prefix
آسبح	آ	اسبح	أ
تسبح/يسبح	ت/ي	تسبح/يسبح	ت/ي
نسبح	ن	نسبح	ن

Table 6. Examples of Slang Suffix for Present verb and their equivalent classical ones

Slang	Present Suffix	Classical	Present Suffix
تلعبين	ين	تلعبي	ي
تلعبون	ون	تلعبوا	ان/ون/ن

Table 7. Examples of Slang Suffix and Prefix for Imperative Verb and their equivalent classical ones

Slang	prefix	suffix	classical	prefix	suffix
اشرب	ا	-	اشرب	أ	-
شربي	-	ي	اشربي	أ	ي
شربوا/شربو	-	وا/و	اشربوا/ اشربا/اشربن	أ	وا/ا/ن

There is no prefix of past tense in Bahraini slang language similarto that in classical language.

Table 8. Examples of Slang Suffix for Past verb and their equivalent classical ones

Slang	Suffix	Classical	Suffix
كتبت	ت	كتبت	ت
كتبتي	تي	كتبت	ت
كتبنا	نا	كتبنا	نا
كتبو/كتبوا	وا/و	كتبوا/كتبا/كتبتن/كتبتن	ن/تن/وا/ا

3.6.2 Nouns Mapping

Most names are constant; they do not need to be converted. However, the help of the look up table must convert some Bahraini slang words.

Examples for Mapping

1- I/q: "ليش يتحرك الحبل بروحه".

2- Determination type of terms:

Table 9. Each Terms Type

1st term	2nd term	3rd term	4th term
Q	V2	N	N

1- Check the validation:

In this case its true format ⟶ Go to the next step

2- Search in two ways:

- The first way search in classical documents:

  a- By checking all the terms in look up table, we can find that (بروحه) in slang has the synonym meaning in the classical which is (لوحده), and we have the question term (ليش؟) that has the same meaning of (لماذا؟).

  b- Now, correct the entire query;.لماذا يتحرك الحبل لوحده

  Here we have a verb (يتحرك) that has a prefix (ي), it will still the same because in present prefix classical the same form (ي). Then the word still (يتحرك). In addition, (حبل) still the same without changing (حبل). Now we find that we got the classical query:؟لماذا يتحرك الحبل لوحده .

  c- Remove the stop words like (لماذا).

  d- Now search in the classical documents.

- Second way; search in the slang documents:

  a- Keep the query and just remove the stop words (ليش).

b- Now search in the slang documents about (يتحرك الحبل بروحه).

Now, show all the result in both the classical and the slang

4. Results and Decision of Results

C# language was mainly used to build this system with the use of Microsoft visual studio of 2013. The searching is (content the dataset) includes a group-stored documents in both slang and classical Arabic language. The process of building the SQL of Microsoft server 2012 needs two main stores. The first, it has around 250 of documents, which have been written in Arabic slang. Not only that, but also the main used vocabularies in it are the spoken ones. They are taken from some sources of the slang document and from the forums too. In addition, the second has a number of classical documents that were record about 493 which are taken either from the social data or from articles. Two stop words also add an important role in this process where they are used with 445 classical ones. Moreover, the list was collected from a social program like marathon, but the second one was taken from a group of collected slang stop words that its numbers are around eighty words.

For both slang and classical, we collected many documents especially from the social media like; Facebook, twitter, Instagram etc. The document is collected manually. Therefore, we faced many difficulties because of the lack of documents in slang language. To solve this problem, we converted some classical documents into slang ones according to our knowledge of Bahraini slang since Bahrain is considered as ourhometown. In addition, we asked for help from Mr. Mohammad Al-bitar that is a specialist in Arabic language, and he is an Arabic teacher.

As a new system, it is very significant to highlight the performance of it. By comparing it with other previous system, we can test it perfectly. However, the main difficulty in this system is that it has a lack of slang documents, which we need. In this research, we listed here some basic measure performances, which are mainly used in the information retrieval system (Precision, Recall and F-measure).

4.1 The Experimental Results

The efficiency of this new system can be known by making both processes:

1- By making comparison between both results of two languages the slang and the classical queries, and by applying it on classical, slang, and the hybrid dataset's content too.

2- By making a second comparison between the values of the precision in slang queries with the other twenty values of precision retrieving documents. Finally, by applying it on Google and Yahoo Search Engine.

4.2 Testing Using Bahraini Slang Language-based over Bahraini Slang Dataset

The system has been tested using Bahraini Slang Language-based Arabic queries that used to retrieve slang documents. Table (10) shows a list of fifty slang-based queries and the precision, recall and F- measure values are shown.

Table 10. Precision, Recall and F- measure values after testing Bahraini Slang Language-based over Bahraini Slang Dataset

Number	The Query	Precision	Recall	F - measure
1	شنو اهي اضرار الدخان ؟	0.67	0.65	0.66
2	جم عدد محافظات الاردن ؟	0.60	0.58	0.59
3	شلون نحل الزحمة ؟	0.58	0.56	0.57
4	شلون انزل وزني؟	0.77	0.74	0.75
5	شنو احسن جامعه في البحرين ؟	0.54	0.56	0.55
6	شنو اسباب البطالة ؟	0.57	0.59	0.58
7	شنو اسباب الطلاق ؟	0.66	0.68	0.67
8	شنو اسباب الكحة ؟	0.67	0.69	0.68
9	ابغي نوع لحاف مريح .	0.79	0.76	0.77
10	ابغي احلى تصميم نفانيف .	0.63	0.66	0.64
11	منو اللي اخترع الكهربة ؟	0.69	0.71	0.70
12	شنو التويتر ؟	0.73	0.77	0.75
13	ليش التلفون يطفي بروحه ؟	0.45	0.54	0.49
14	ليش الموتر ما يشتغل ؟	0.56	0.55	0.55
15	الاغراض اللي يحتاجها اليهال .	0.68	0.78	.730
16	انواع الفواكه والخضار .	0.57	0.67	.620

17	ياريت الزمان يرجع .	0.35	0.40	0.37
18	نكت تضحك .	0.57	0.68	0.62
19	صلطة الفواكه .	0.75	0.72	0.73
20	الزهيوي من الحشرات ؟	0.80	0.65	0.72
21	هدية عشان عرس .	0.33	0.50	0.40
22	الحيوانات اللي تعيش في البيت .	0.84	0.55	0.67
23	حلى مشهور في البحرين .	0.65	0.62	0.63
24	اكلة مشهوره في البحرين .	0.50	0.75	0.60
25	تركيا من احلى الاماكن السياحية .	0.51	0.72	0.60
26	ارتفاع اسعار اللحم في البحرين .	0.40	0.64	0.49
27	اكثر الشوارع زحمة في البحرين .	0.70	0.69	0.69
28	اكثر مرض منتشر عند العيايز .	0.72	0.76	0.74
29	شنو الحل انباق حسابي ؟	0.59	0.64	0.61
30	وين يطلع الفقع ؟	0.47	0.20	0.28
31	من اطول ريال بالعالم ؟	0.36	0.60	0.45
32	شنو استخدامات التلفون ؟	0.60	0.69	0.64
33	جم عدد الكواكب ؟	0.78	0.53	0.73
34	متى تبدا عطلة الربيع ؟	0.52	0.60	0.56
35	شلون تسوين الكيك ؟	0.76	0.53	0.62
36	وين تعيش الارانب ؟	0.62	0.55	0.58
37	شلون اسوي بحث ؟	0.74	0.75	0.74
38	شلون تطبخين السمج ؟	0.81	0.70	0.75
39	شلون انزل فيديوهات ؟	0.52	0.65	0.57
40	ليش نشرب العصير ؟	0.59	0.63	0.61
41	شنو اخر فيلم بالسينما ؟	0.64	0.76	0.69
42	الاغراض اللي تحتاجها العروس ؟	0.53	0.65	0.58
43	البس عشان ما تبرد ؟	0.54	0.75	0.63
44	منو اكثر ناس تشيش ؟	0.43	0.66	0.52
45	شنو تبي تدرس بالجامعه ؟	0.45	0.71	0.55
46	جم عدد القارات ؟	0.77	0.61	0.68
47	اكثر دوله تزرع زيتون .	0.66	0.70	0.68
48	شلون تخيطين النفنوف ؟	0.70	0.60	0.84
49	باي قاره البحرين ؟	0.81	0.64	0.72
50	منو بنات الرسول ؟	0.83	0.63	0.72

4.3 Testing Using Arabic Classical Language-based over A Classical Dataset

Table (11) shows the precision, recall and F- measure values when testing a classical dataset but by using the equivalent classical for the queries used in Table (10).

Table 11. Precision, Recall and F-measure values after testing a classical dataset using classical-based Arabic queries

Number	The Query	Precision	Recall	F - measure
1	ما هي اضرار الدخان ؟	0.65	0.56	0.60
2	كم عدد محافظات البحرين ؟	0.60	0.65	0.62
3	كيف نحل ازمة السير ؟	0.68	0.64	0.66
4	كيف انزل وزني؟	0.77	0.70	0.73
5	ما هي افضل جامعه بالبحرين ؟	0.64	0.71	0.67
6	ما اسباب البطالة ؟	0.57	0.66	0.61
7	ما اسباب الطلاق ؟	0.66	0.75	0.70
8	ما اسباب الكحة ؟	0.67	0.70	0.68
9	ادرسوا كي تنجحوا .	0.69	0.40	0.51
10	اريد اجمل تصميم فساتين .	0.61	0.70	0.65
11	من الذي اخترع الكهرباء ؟	0.69	0.57	0.62
12	ما هو التويتر ؟	0.73	0.68	0.70

13	لماذا التلفون يغلق ؟	0.51	0.78	0.62
14	لماذا السياره لا تشتغل ؟	0.60	0.60	0.60
15	الاغراض التي يحتاجها الاولاد .	0.68	0.58	.630
16	انواع الفواكه والخضار .	0.57	0.67	.610
17	ياليت الزمان يعود .	0.43	0.35	0.34
18	نكت تضحك .	0.57	0.66	0.61
19	سلطة الفواكه .	0.77	0.55	0.64
20	الصرصور من الحشرات ؟	0.82	0.40	0.54
21	هدية لزواج .	0.49	0.50	0.49
22	الحيوانات التي تعيش في البيت .	0.84	0.53	0.65
23	حلويات مشهورة بالبحرين .	0.81	0.60	0.69
24	طبق مشهور بالبحرين .	0.78	0.57	0.66
25	تركيا من اجمل المناطق السياحية .	0.55	0.60	0.57
26	ارتفاع اسعار اللحوم بالبحرين .	0.35	0.61	0.44
27	اكثر الشوارع ازمة بالبحرين .	0.68	0.60	0.64
28	اكثر مرض منتشر عند كبار السن .	0.78	0.57	0.66
29	ما الحل انسرق حسابي ؟	0.56	0.61	0.58
30	اين يطلع الفقع ؟	0.51	0.20	0.29
31	من اطول رجل بالعالم ؟	0.31	0.58	0.40
32	ما استخدامات التلفون ؟	0.79	0.70	0.74
33	كم عدد الكواكب ؟	0.84	0.55	0.66
34	متى تبدا عطلة الربيع ؟	0.80	0.57	0.67
35	كيف تعملي الكيك ؟	0.76	0.49	0.60
36	اين تعيش الارانب ؟	0.51	0.60	0.55
37	كيف اعمل بحث ؟	0.74	0.60	0.66
38	كيف بطبخ سمك ؟	0.52	0.75	0.61
39	كيف بنزل فيديوهات ؟	0.49	0.62	0.55
40	لماذا نشرب العصير ؟	0.47	0.68	0.56
41	ما اخر فيلم بالسينما ؟	0.70	0.70	0.70
42	الاغراض اللي بتحتاجها العروس ؟	0.75	0.53	0.62
43	البس كي لا تبرد ؟	0.59	0.63	0.61
44	من اكثر ناس بتارجل ؟	0.50	0.61	0.55
45	ماذا تريد تدرس بالجامعه ؟	0.45	0.73	0.56
46	كم عدد القارات ؟	0.62	0.56	0.59
47	اكثر دوله تزرع زيتون .	0.53	0.66	0.59
48	كيف تخيطين فستان ؟	0.72	0.59	0.65
49	بأي قاره البحرين ؟	0.63	0.61	0.62
50	من بنات الرسول ؟	0.83	0.63	0.72

Figure (8) compares the average values for precision, recall and F- measure. It shows that the results of the recall precision of the slang are very close to the ones of the classical precision because in our system we put special rules of slang. In addition, we have put specific rules to make the inversion process from the slang form to the classical one. Therefore, the precision results were about 0.64 for both researches. However, the result of the recall in our researching system in both slang and classical are 0.62 and 0.60 respectively. These are good rates in retrieving the relevant documents and they are considered as good evaluation for the system.

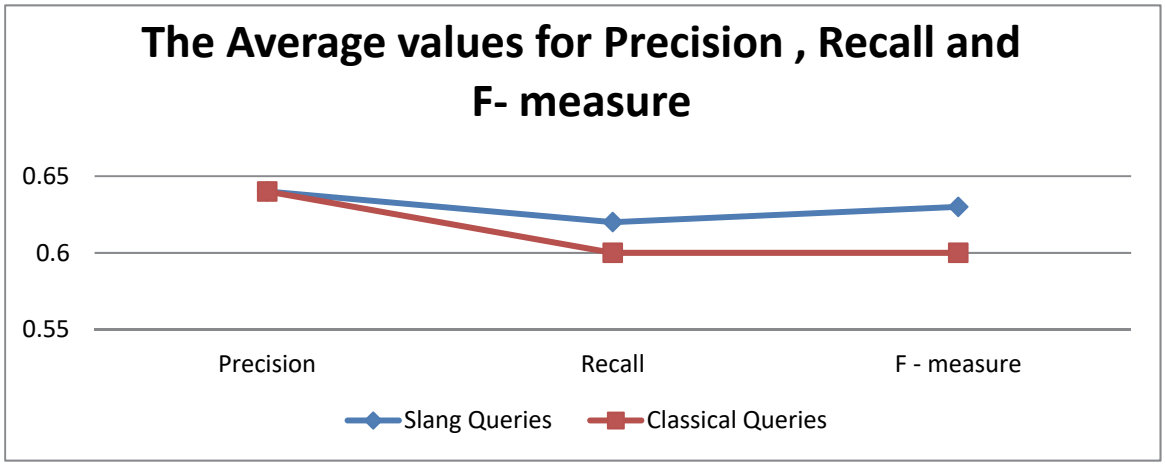

Figure 8. The average values of precision, recall and F- measure values after testing for both slang and classical queries

4.3.3 Comparing Research System with Google and Yahoo Search Engine

To make my point clear and to prove the efficiency of this system, we applied the group of queries, the slang ones that were mentioned in Table (10) on both previous engines. Thus, the first twenty retrieved results were calculated in both. The comparison between results was obtained in the Research system by applying the slang queries. Which helped this comparison to be done is that both engines; Yahoo and Google have documents of both kinds of slang and classical. In Table (12), the resulted values of precision in both engines are mentioned. To make everything clear, we used precision in this comparison because of its high possibility of counting its relevant results after applying the queries. However, we must confess that the recall can never be counted since the number of its relevant documents that are stored in their databases are unknown.

In this step, we searched about each query in both Google and yahoo, and then we found the first 20 results about each one. Then, we figured out whether each one is relevant or irrelevant. After that, we divided the relevant document on all the retrieved ones where the result was a precision value for each query.

Table 12. Precision values given by Research System, Google and Yahoo Search Engine

Query No.	Research System	Google	Yahoo
1	0.67	0.68	0.40
2	0.60	0.50	0.20
3	0.58	0.40	0.10
4	0.77	0.52	0.20
5	0.54	0.40	0.10
6	0.57	0.60	0.35
7	0.66	0.52	0.28
8	0.67	0.49	0.35
9	0.79	0.44	0.20
10	0.63	0.40	0.10
11	0.69	0.75	0.50
12	0.73	0.70	0.45
13	0.45	0.30	0.10
14	0.56	0.33	0.10
15	0.68	0.40	0.20
16	0.57	0.60	0.25
17	0.35	0.20	0.00
18	0.57	0.75	0.40
19	0.75	0.73	0.50
20	0.80	0.53	0.20
21	0.33	0.00	0.00

22	0.84	0.60	0.30
23	0.65	0.35	0.00
24	0.50	0.43	0.10
25	0.51	0.55	0.20
26	0.40	0.30	0.00
27	0.70	0.55	0.15
28	0.72	0.00	0.00
29	0.59	0.10	0.00
30	0.47	0.50	0.10
31	0.36	0.45	0.00
32	0.60	0.50	0.20
33	0.78	0.80	0.50
34	0.52	0.55	0.15
35	0.76	0.60	0.20
36	0.62	0.70	0.40
37	0.74	0.50	0.10
38	0.81	0.70	0.35
39	0.52	0.40	0.10
40	0.59	0.51	0.30
41	0.64	0.34	0.10
42	0.53	0.25	0.00
43	0.54	0.15	0.00
44	0.43	0.10	0.00
45	0.45	0.28	0.00
46	0.77	0.80	0.45
47	0.66	0.70	0.30
48	0.70	0.65	0.30
49	0.81	0.70	0.30
50	0.83	0.85	0.40

Figure (9), Reveals a clear comparing in the precision has resulted averages that were given in Table (12). As a result, we can notice how the Research system shows better values when it comes to precision, and they are even highly better than other searching engines like Google and yahoo. Although both of them support the process of searching in Arabic, but no one can deny that the results in the Research system are better than theirs especially in the use of written slang queries.

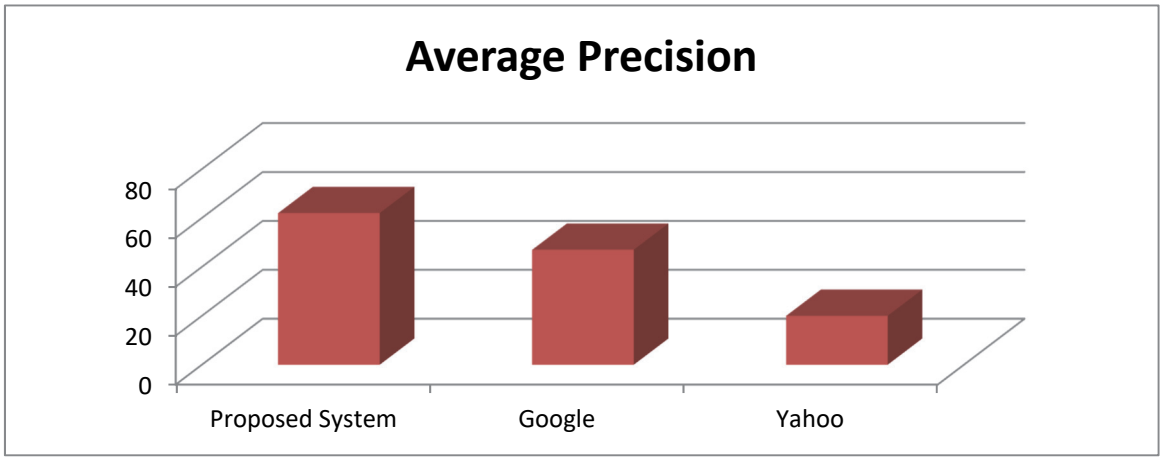

Figure 9. Comparison Results between the average precision values given by Research System, Google and Yahoo Search Engine

5. Conclusion and Future Work

5.1 Conclusion

Classical documents results were all close in three stages of: precision, recalling, and f-measuring either in slang or classical. This is justifiable because the Research conversion process helps in converting the slang concepts to the equivalent words in classical. Not only that, but also the slang document was used in searching for both slang

This research has developed the Arabic information retrieval process (IR) through providing a new searching system using Bahraini slang queries. By using the Research Arabic slang grammar (ASG), results reveal that high accuracy was achieved in the retrieved documents while helping the user's query to retrieve classical ones. In addition, Arabic classical documents results were all close in three stages of: precision, recalling, and f-measuring either in slang or classical. This is justifiable because the Research conversion process helps in converting the slang concepts to the equivalent words in classical. Not only that, but also the slang document was used in searching for both slang and classical queries. The results reveal that using the slang-based queries dominate the use of classical ones but not the opposite. Moreover, results also show that the use of slang-based queries is better than the use of classical ones in the retrieving document.

We can notice that both queries can be used for searching, and both kinds of quires can be input. Overall, results also proved that using slang queries is extremely better than using classical.

5.2 Future Work

We hope in the future to expand the Research system on other countries languages and to categorize the field of searching for instance: politics, geography, etc. In addition, some researcher in the future can working hard to find a way to differentiate between typing errors and the wanted concepts, supporting the system by multimedia types like pictures, videos, voices etc, and helping disable people in the process of searching by adding a new technique of searching for instance; by voice recording especially for blinds and handy loss.Thus, it will be very easy since slang and accents are better noticed in speaking not just in writing.

References

Abuata, B., & Al-Omari, A. (2014). A Rule-Based Stemmer for Arabic Gulf Dialect. *Journal of King Saud University– Science*. https://doi.org/10.1016/j.jksuci.2014.04.003

Alamlahi, Y., & Ahmed, F. (2011). Sana'ani Dialect to Modern Standard Arabic: Rule-based Direct Machine Translation. In: proceeding of the *2011 International conference on artificial intelligence (ICAI'11)*.

Albalooshi, N., Mohamed, N., & Al-Jarood, J. (2011). The challenges of Arabic language use on the Internet. *International Conference for Internet Technology and Secured Transactions*, ICITST 2011.

Al-Gaphari, G., & Al-Yadoumi, M. (2010). A method to convert Sana'ani accent to modern standard Arabic. *Int. J. Inf. Sci. Manage, 8*(1), 39–49.

Almeman, K., & Lee, M. (2013). Automatic building of Arabic multi dialect text corpora by bootstrapping dialect words. Communications, Signal Processing and their Applications *(ICCSPA), 1st International Conference* on 12–14 Feb, pp. 1–6.

Al-Shalabi, R., Kanaan, G., & Al-Sarayreh, B. (2009). Proper noun extracting algorithm for Arabic language. In: Proceedings of the *International conference on information technology to celebrate* S. Charmonman's 72nd birthday. 30 March; Thailand. pp. 28.1–28.9.

AL-Shalabi, R., Kannan, G., & AI-Serhan, H. (2003). New approach for extracting Arabic roots. Proc. of 2003 *International Arab conference on Information Technology (ACIT'2003), Alexandria*, pp. 42-59.

Azmi, A., & Aljafari, E. (2015). Arabic tweets sentiment analysis - A hybrid scheme. *Journal of Information Science.* Retrieved from sagepub.co.uk/journalsPermissions.nav.

Azmi, A., & Aljafari, E. (2017). Performance evaluation of different age groups for gestural interaction: a case study with Microsoft Kinect and Leap Motion. *Univ Access Inf Soc., 17,* 131. https://doi.org/10.1007/s10209-017-0522-3

Diab, M., Habash, N., Rambow, O., Altantawy, M., & Benajiba, Y. (2010). COLABA: Arabic Dialect Annotation and Processing. In: proceeding of the *workshop on Semitic language processing (LREC-2010)*, PP.66-74.

Duwairi, R., Al-Refai, M., & Khasawneh, N. (2007). Stemming versus light stemming as feature selection techniques for Arabic text categorization. *Innovations in Information Technology, IIT 07. 4th International Conference* on, 18–20 Nov, pp. 446–450.

Elmasry, M., Soliman, T., & Hedar, A. (2014). Sentiment Analysis of Arabic Slang Comments on Facebook. *International Journal of Computers and Technology (IJCT), 12,* 3470-3478. http://dx.doi.org/10.24297/ijct.v12i5.2917

Ghoneim, M. (2010). Arabic Information Retrieval Systems: Aspects of ambiguity and the potential solutions [electronic study], Saudi Arabia. Retrieved August, 2018, from http://www.kfnl.gov.sa/idarat/alnsher%20el/nothom/PubMain.htm

Gomaa, Y. (2015). Saudi Youth Slang Innovations: A Sociolinguistic Approach. *International Journal of Linguistics and Communication*, *3*(2), 98-112. http://dx.doi.org/10.15640/ijlc.v3n2a10

Goweder, A., Alhami, H., Rashed, T., & Al-Musrati, A. (2008). A hybrid method for stemming Arabic text. In *International Arab conference on information technology (ACIT)* 16–18 December; Hammamet City, Tunisia.

Harrag, F., El-Qawasmah, E., & Al-Salman, A. (2011). Stemming as a feature reduction technique for Arabic text categorization. Programming and Systems (ISPS), *10th International Symposium* on, 25–27 April, pp. 128–133.

Kadri, Y., & Nie, J. (2006). Effective stemming for Arabic information retrieval. In Proceedings of *the Challenge of Arabic for NLP/MT Conference*. The British Computer Society. London, UK.

Mutahhar, A., & Watson, J. (2002). Social issues in popular Yemeni culture. *Yemeni– British project supported by the British Embassy, Social Fund for Development and Leigh Douglas Memorial Fund, Sana'a, Yemen.*

Ogheneovo, E., & Japheth, R. (2016). Application of Vector Space Model to Query Ranking and Information Retrieval*. International Journal of Advanced Research in Computer Science and Software Engineering Research Paper, 6*(5). Retrieved from www.ijarcsse.com

Rogerson, E. (2008). An evaluation of existing light stemming algorithms for Arabic keyword searches. Master Thesis. *The University of North Carolina*.

Sanan, M., Rammal, M., & Zreik, K. (2008). Internet Arabic search engines studies. In: Proceedings of *Information and communication technologies: From theory to applications ICTTA*; 7–11 April; Damascus, Syria. pp. 1–8.

Soliman, T., & Hedar, A. (2014). MINING SOCIAL NETWORKS' ARABIC SLANG COMMENTS. *In Proceedings of IADIS European Conference on Data Mining 2013 (ECDM'13),* 22 – 24 July, Prague, Czech Republic.

Wijewickrema, P., & Ratnayake, A. (2013). Enhancing Accuracy of a Search Output: A Conceptual Model for Information Retrieval. *Journal of the University Librarians Association of Sri Lanka, 17*(2), 119-135.