

Enhanced Arabic Information Retrieval by Using Arabic Slang Language

Mustafa Abdel-Kareem Ababneh¹, Ghassan Kanaan¹ & Ayat Amin Al-Jarrah¹

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

Correspondence: Mustafa Abdel-Kareem Ababneh, Amman Arab University, Amman, Jordan. E-mail: mus2010@yahoo.com

Received: April 1, 2019

Accepted: April 28, 2019

Online Published: May 23, 2019

doi:10.5539/mas.v13n6p24

URL: <https://doi.org/10.5539/mas.v13n6p24>

Abstract

Slang language has become the most used language in the most countries. It has almost become the first language in the social media, websites and daily conversations. Moreover, it has become used in many conferences to clarify information and to deliver the required purpose of them. Therefore, this great spread of slang language over the world. In Jordan indicates that it is important to know meanings of Jordanian slang vocabularies. Mainly, In research system, we created a system framework allows users to restore Arabic information depending on queries that are written in slang language and this framework was made basically by context-free grammar to convert from slang to classical and vice versa. In addition, to conclude with, we will apply it on the colloquial slang in North of Jordan specifically; Irbid, Ajloun, Jerash, Mafraq and AlRamtha city. As well as, we will make a special file for Non_Arabic words and the stop words too. After we made an evaluation for the system relying on the results of recall, precision and F-measure where the results of precision about 0.63 for both researches slang and classical query, and this indicates that the system supports searching in Jordanian slang language. The purpose of this research is to enhance Arabic information retrieval, and it will be a significant resource for researchers who are interested in slang languages. As well as, it helps tie communities together.

Keywords: jordanian arabic slang language grammar, slang language, classical language

1. Introduction

1.1 Introduce the Problem

Nowadays, the slang languages usage is one of the most popular languages in many social, academic, political and scientific majors. Everyday conversations (Almeman & Lee, 2013), messages, letters, and social media (Harrag and et al., 2011; Duwairi and et al., 2007; Al-Shalabi and et al., 2003), they are originally taken from standard and they are used in slang language. There are twelve Arabic dialects, which are used in 28 countries, but the most common one between all these dialects is the Egyptian one.

Arabic language is an old language precedes Islam (Azmi, 2017). Many Arabic concepts are missed in the new born generations who have lack of knowledge toward some terms and their meanings. Therefore, as a result, these generations do not use those terms at all. Some researches indicated that there are a lot of increasing in the native speakers in the classical language of Arabic (MSA) (Mutahhar & Waston, 2002).

It is very noticeable that Information Retrieval Science has many studies that help users in restoring the important information only (Ghoniem, 2008; Sanan and et al., 2008).

It is noticeable also that each Arabic dialect is different from another according to its area. Therefore, every vocabulary interfere with MSA around 80% or more. Therefore, those different aspects include very popular words like see, not, and go. Simultaneously, they have phonology, morphology, and syntax rules (Habash & Rambow, 2006).

However, its known that have many properties which distinguish the Arabic language from other ones in many ways like writing's direction; if it's left or right (Benajiba and et al., 2007; Nwesri, 2008). These properties can be changeable according to the position in each statement (Ghoneim, 2008; Sanan and et al., 2008; Al-Shalabi and et al., 2009). Eventually, we should acknowledge that many algorithms can take the root of any word and search about it in dictionaries (Goweder and et al., 2008; Al-shalabi and et al., 2005, Ahmed and et al., 2008), but unfortunately up to now there is no algorithm that can completely find any stem of Arabic words (Goweder and et

al., 2008).

1.2 Research Problem

Mainly, the research problem is that there is no engine for using Jordanian slang query.

1.3 Problem Solution

Thus, we worked hard to uphold the system of information's restoring that includes the great value of our study.

In addition, our retrieving information's system is used by depending on the colloquial not the standard that is now a very valuable tool in evaluating the engine's search. As a result, in this algorithm we tried to help all ages and levels by making the search through the internet as easy as possible, for example: young, old, students, professors, ignorant, knowledgeable, Arabs, non-Arabs, etc.

2. Related Studies

It was suggested by many researches that when Arabic texts indexing uses stems this will increase the retrieval effect above all the words and stems. As a result, most algorithms were highly improved to help in extracting Arabic words stems.

It is a significant issue to supply relevant information for users, which saves both time, and effort for whole sides of the process. Eventually, the improved models of IR by presenting them through computers as numerical models will be a high step in the modernity and efficiency of IR gadgets. Some study introduces a new way of IR that at first appears as a vector of different distances between both the concepts of the used queries and the distances of the same data or documents concepts. As a result, it distinguishes the relevant results in queries and documents by matching distances of each two terms in the query with the distances among the similar terms, in every corpus documents. In addition, this research is considered as an example to show the ability of the new model in showing the two or more documents depending on the given query that really cannot be done by only using the vector space model as the documents include the similar frequent concepts and the same number of whole concepts (Wijewickrema & Ratnayake, 2013).

Information Retrieving is still a very problematic issue because it is considered as a time wasting especially if it is used from internet and long data. Thus, many algorithms and strategies were made to solve this issue and to make it easy. However, retrieving is continuing to be very sophisticated. Many applications were used to add, delete, and change like the Book Library System, and the Commercial Document Retrieval Services. This constrains the types of data and algorithms that are built to be used for the IR. In some study, the strategy of VSM (Vector Space Model) of IR is mainly used. Firstly, we find the percentages of similarities in scores usage for each item and we compare the results. Then, the cosine measure is used in order to compute the similarity measure and to check the angle between the documents vector and the query vector where VSMs are mainly based on geometry and every term has its own dimension in a multiple space of dimensions, queries, and documents. Eventually, we found that it is easier to retrieve the data or the information based on the similarity and to produce an efficient strategy for IR (Ogheneovo & Japheth, 2016).

2.1 Arabic Information Retrieval

In (Al Kharashi & Al Sughaiyer, 2004), the used method can professionally define groups of rules and apply them on the Arabic terms to find their stems easily without the need for a complicated process or computations.

Dialectal Arabic Information Retrieval Assistant (DIRA) is defined as a wide query method that gives a search term of Arabic Classical in general or its dialects, it is even provided of English too. Recently, the retrieval of Arabic text became very needed according to the rise of dialectal content especially of social media. DIRA shows the difficulties of retrieving information in Arabic dialects that have a big role in the languages differences from classical one. In addition, we must acknowledge that DIRA is the only way to generate the search of dialect terms with relevant various linguistic forms of both English and standard Arabic (Diab and et al., 2013).

2.2 Algorithm in Gulf Slang Language

The Gulf language is a very popular one in many countries. Moreover, after having obsolete words of the Gulf dialects, many algorithm were used to extract the root of slang words by deleting the affixes in words and comparing it with other non-Arabic words with stop terms and by assuming that the root has three letters. As a result, after applying this algorithm on the Gulf dialects, the levels of accuracy of the extracted roots were very high even higher than all the other roots with a few of affixes and they became one of the characteristics of Gulf speakers (Abuata & Al-Omari, 2015).

Another research tried to plan a very specific system of information retrieval, which reveals our attention against

the background of one of the local dialects in Saudi Arabia. Therefore, in this system, they depend on adapting an MSA rooting strategy and a group of slang MSA diversion rules that are based on dictionary (Azmi, 2015).

2.3 Algorithm in Jordanian Slang Language

Other studies also were concerned with building a framework in order to enhance the retrieval accuracy of Arabic information that use slang query and gives the system a way of dealing with query. So, this algorithm is based on context free grammar "CFG" that neglected the root of past, present, and command verbs in three steps which are: ignoring nouns, making sure if the syntactic rules for sentences are correct or not, and stripping their affixes (Shatnawi and et al., 2012).

2.4 Algorithm in Yemeni Slang Language

As we noticed, there is an increase of using the slang language in many countries such as the concern of researchers about the Yemeni language "Sana'a language" to improve it to a very modern standard one. This algorithm is based on deleting vowels, suffixes, and prefixes. After that, the authors take the root and compare it with stems of slang Yemeni one and MSA and it finds the texts in MSA (Modern Standard Arabic). The result of this algorithm is 16.29% of these concepts, which has distorted suffix, 0.70%, contain prefix, and 2.16% contains stem. (Al-Gaphari & Al-Yadoumi, 2006).

2.5 Retrieving in Social Networks

Activities in the social media and network became very noticeable and the slang one is commonly used, that is why many researches were done and one of them mainly uses the SSWIL (slang sentimental words and lexicon) of various words. Gaussain Kernel's SVM (Support Vector Machine) is used as a tool to classify the social media comments such as Facebook that are written in Arabic. Many statistics were used to reach a very honest and precise percentage, and the results were very high about more than 88% for precision and more than 75% for the recall (Elmasry and et al., 2014).

Another research done to analyze young's comments on Facebook we need to study the slang Arabic that they use to communicate with each other in everyday conversations. That is why we chose the SVM classifier to apply our study. In addition, it must be known that the SVM includes three main aspects: first the preparation of Arabic comments (data), and both data of division and of preprocessing too. Not just for Facebook, these studies also took care of Twitter and other topics. The rates were also satisfying because of the use of SSWIL where 75% of comments were replaced by 86% comments. Thus, they were recorded as: more than 88% for precision instead of 82%, and 78% of recall instead of 59% (Soliman & Hedar, 2014).

Also, another study as a sociolinguistic one because it discusses the young Saudi slang. Unlike the other cities, the research in Saudi Arabia was very exhausted to find results about this linguistic issue by focusing on the social factors, the source of acquisition, and the reasons of use too. In addition, it concentrates on the common topics that this slang is interested about.

One more noticeable thing is that the relationship between age and the slang is negative; the younger the person is, the more input he/she has. While talking about gender, it's clear that males use slang more than females which are obvious because both genders have different interests while using words, concepts, and even topics especially while talking about sexual ones (Gomaa and et al., 2015).

3. Methodology

This research aims to design framework that works to restore the Arabic language information by using the Jordanian slang language, whereas the processes of this system are based on a group of rules that should be done before restoring the information. Moreover, this system works to normalize the query that is written in slang, then it searches for it in the engine, after that it restores a group of documents in the classical form and the slang one. In addition, this document was collected from the social media that contains a huge combination of slang document.

3.1 The System Framework

Figure (1) represents the framework of the Arabic retrieval system by using the North Jordanian slang query. It shows the most important steps that the system should follow to restore the information by using the Jordanian slang for the North of Jordan.

This system works by checking the sentences grammatical structure by applying the special rules of this system. If the structure is wrong, it asks the user to rewrite the query again. However, if the arrangement is true, it goes to the next step where the system is divided into two branches.

The first one converts the query from slang form into the classical one by applying the assigned rules. Then, it

searches for the terms inside the two lists of non-Arabic list and the look up tables. In addition, it searches about the resulted query inside the classical document. The second branch leaves the query on its shape in the slang, and it deletes the stop words, instead. It also searches for the unknown words inside the non-Arabic list and the look up table. After that, it searches about it in the slang document. Finally, it restores all the resulted documents for the user if it's either slang or classical.

3.2 Jordanian Arabic Slang Language Grammar (JASG)

The system of restoring the slang language is based mainly on a large number of rules (JASG) where the research for the query depends on the rules. If it does not match the rules, the user will be asked to write a new query.

In this stage, the type of words in the queries must be determined if it is noun, verb, question mark, or preposition...etc.

3.3 Definition of the Grammar

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

- Noun clause (Ns): the most significant part is the noun (N).
- The verbal clause: the most significant part is the verb (V).
- Interrogative sentence: the most significant part is the question mark (Q).

The content of the grammars of the Jordanian Slang language is shown in Figure (2), and the list of the used grammars is shown in Table (1).

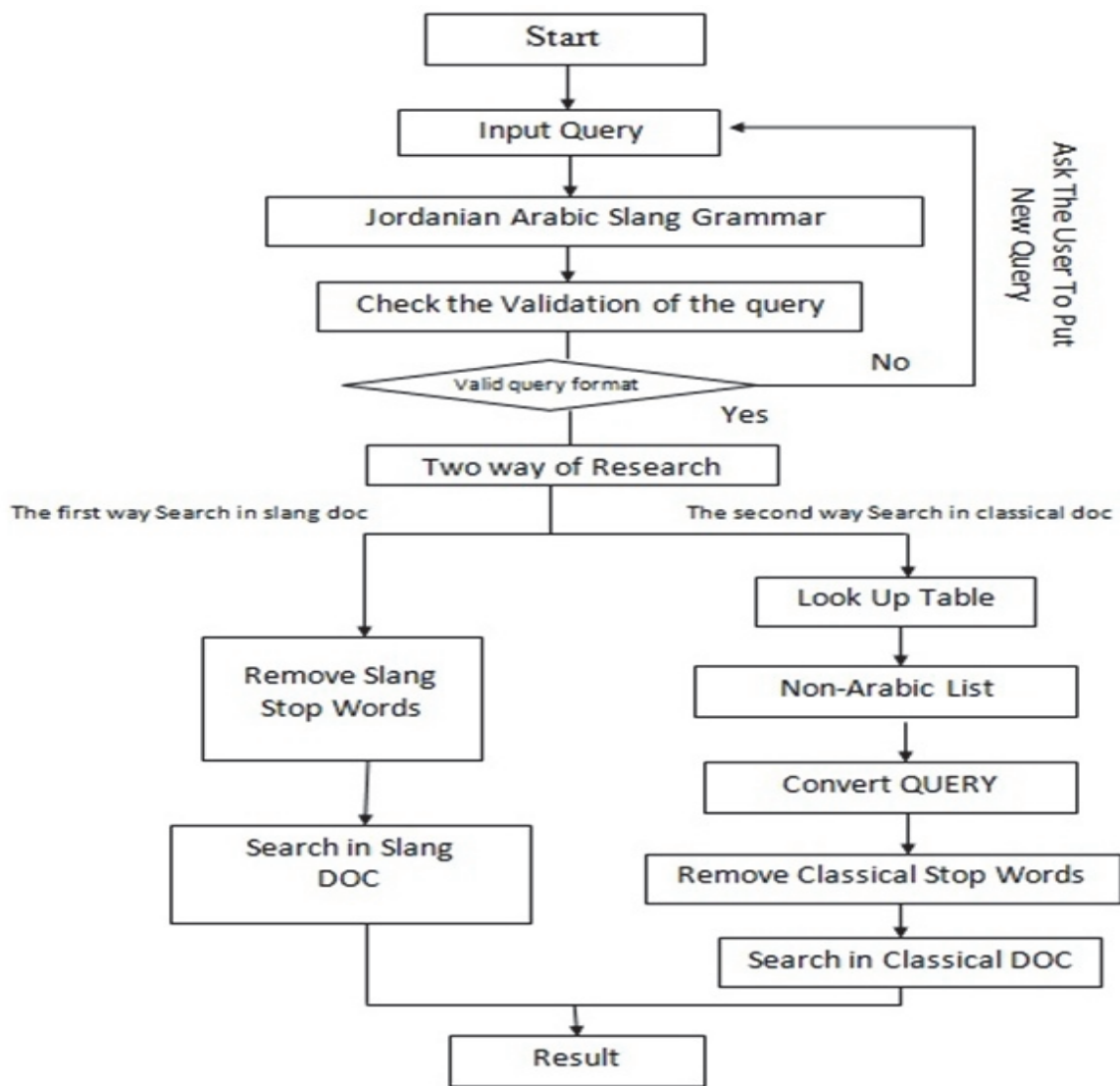


Figure 1. Research Framework

Table 1. Shortcuts of Grammar

Abbreviation	Meaning
Pre1V	List of prefix for present verb {بت ب بن بي ت }
Pre2V	List of prefix for imperative verb { }
Suf 1V	List of suffix for present verbs { ن وا ي }
Suf2V	List of suffix for past verbs { ت تي توا تن ن ي }
Suf3V	List of suffix for Imperative verbs { ن وا ي }
Pre N	List of prefix letter of nouns {ال لل وال بال ول }
Suf N	List of suffix letter of nouns {ات ان ابن ون ك ج كم و نه ا تيه ا تني }
Number table (No)	ثلثعش { عشرة احدعش اثنعش {واحد اثنين ثلاثة اربعة خمسة ستة سبعة ثمانية تسعة { اربععش خمسعش ستعش سبععش ثمنعش تسععش عشرين
Jar table (Jar)	{من الى عن على مع ب في فال ل عال }

Calling table (Cal)	{ يا }
Zarf table (Zarf)	{ عند جوا خلف ورا جنب قبل بعد تحت بين قدام فوق شهر يوم دقيقة لحظة ثانية }
Mawsool names table (MS)	{ اللي الي }
Jazem table (JZ)	{ ما لا }
Naseb table (NS)	{ عشان }
Question (ASK)	Question terms { مين شلون لمين وين اي كيف ايمنه ليش شو ليه ايش جيف كيف قديش } {باي متى }
SufASK	Suffix of question { كم ن ك ج }
SN	{ هاظ هاي هذول هناك غاد }
Kan	{ كان صار ظل بعده مدام مش راح كان رح }
Pro	{ انت انتي انا اني همه هن انتوا انتن انتم هو هي همه نحن احنا }
Fn	{ ابو اخو سلفي سلف }
En	{ ان كانه كان كنه لكن بس لو انه ياريت }
Negative (Neg)	{ ما مش مو لا }
Atf	{ و او }

(A)

- 1- **SHO**= {Pre1V, Pre2V, Suf1V, Suf2V, Suf3V, PreN, SufN, No, Jar, Cal, Zarf, MS, JZ, NS, ASK, SufASK, SN, EN, Kan, Pro, FN, V1P, V2P, V3I, W, Neg, Atf }
- 2- **WPreN** = { No, Jar, Cal, Zarf }
- 3- **WPreV** = { MS, JZ, NS }
- 4- **V1P** → V | V Suf2V
- 5- **V2P** → V | Pre1V V | Pre1V V Suf1V
- 6- **V3I** → V | Pre2V V | Pre2V V Suf3V
- 7- **W1** → W | PreN W | PreN W SufN | W SufN
- 8- **ASK1** → ASK1 | ASK SufASK

NPH → Pro W1 V1P WSS | Pro W1 V2P WSS | Pro W1 Neg V1P WSS | Pro W1 Neg V2P WSS | Pro W1 WSS | Pro W1 Atf W1 | W1 V1P WSS | W1 V2P WSS | W1 V3I WSS | W1 Neg V1P WSS | W1 Neg V2P WSS | W1 V1P Atf V1P | W1 V2P Atf V2P | W1 V3I Atf V3I | W1 Neg V1P Atf Neg V1P | W1 Neg V2P Atf Neg V2P | W1 PPI V2P | W1 PPI WSS | SN W1 WSS | SN W1 Atf W1 | SN N V1P WSS | SN V1P Atf V1P | SN V2P Atf V2P | SN V1P WSS | SN V2P WSS | SN Neg V1P WSS | SN Neg V2P WSS | EN W1 WSS | EN SN WSS | EN V2P | W1 WSS | EN W1 V2P | EN W1 Neg V2P | Kan W1 W1 WSS | Kan W1 V2P WSS | Kan V2P W1 WSS | FN W1 | SN FN | FN V2P WSS | Pro V1P | Pro V2P | Pro V3I | Pro Neg V1P | Pro Neg V2P | Pro V1P Atf V1P | Pro V2P Atf V2P | Pro V3I Atf V3I | Pro Neg V1P Atf Neg V1P | Pro Neg V2P Atf Neg V2P | Pro V1P WSS | Pro V2P WSS | Pro V3I WSS | Pro Neg V1P WSS | Pro Neg V2P WSS | Pro MS V1P WSS | Pro MS V2P WSS | Pro MS Neg V1P WSS | Pro MS Neg V2P WSS

VPH → V1P W1 WSS | V2P W1 WSS | Neg V1P W1 WSS | Neg V2P W1 WSS | V1P WSS W1 | V2P WSS W1 | Neg V1P WSS W1 | Neg V2P WSS W1 | PPI WSS | V1P FN W1 PPI | V2P FN W1 PPI | V3I N PPI | PPI V2P

(B)

Figure 2. (a&b) Jordanian Slang Language Grammar

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

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

Symbols	Meaning
SHO	The symbols of grammar are showed in table 3.1
WPreN	The tools that precede nouns
WPreV	The verbs that are preceded by tools
V1P	Past tense verbs with affixes
V2P	Present tense verbs with affixes
V3I	Imperative verbs with affixes
W1	Noun with affixes
ASK1	Tools of question with affixes
WSS	All the structures of nouns
PPI	All the structures of verbs
NPH	phrases of nouns
VPH	Phrases of verbs
QPH	Phrases of question

3.4 How to Determinate the Term Type

In this step, the query sentences structure should be checked to know if it is right or wrong according to groups of rules by comparing the written query with the correct written structures inside the grammar. The sentence should include either verbs, nouns, or question marks, with some stop words that will be deleted later if they were found. Figure (2) shows the structures of the correct sentences; the special grammars for this system. In addition, there are a lot of affixes that are related to the verbs, nouns, and question marks, which were clearly and briefly shown

in Table (1).

3.4.1 How to Determinate the Verb Type

The verb is one of the main parts of sentence structure; it has three types past, either present, or imperative. It has a group of different affixes according to the verbs type: past, present, or imperative. The number of the affixes may be either one letter, two or none. In addition, Figure (3), shows pseudo code of Arabic verbs that is used for distinguishing the verb's type.

The present verb is normally connected with both the suffix and the prefix like the verb: (تدرس). It is clear that it's preceded by a prefix, but it is not followed by a suffix. However, in this example; the verb (يتدرسي) is preceded by a prefix and a suffix.

The same as the imperative verb that may be followed by prefix or a suffix like (ادرسوا) where it is preceded by a prefix. When it comes to the past verb, it is different from the present and the imperative. The past does not relate with any prefix, but it relates with a suffix just like this example (درست).

If the verb contains than four letters or more, it may contains some infixes and these are some cases in the past tense: (استعملت), the present verb (يستعمل), and the imperative verb (استعمل).

The verb also includes some equipments, which are clearly defined in Table (1) that shows relative pronouns, the conditional pronouns, and the particles.

In addition, verbs are preceded by relative pronouns (MS), both sealing tools (JZ) and monument tools (NS), and this section was considered as part of the grammar.

```

If (prefix of term in the list of Pre1V || suffix of term in the list of Suf2V)
Then
Term_type = present verb
Else if (prefix of term in the list of Pre2V || suffix of term in the list of Suf3V)
Then
Term_type = imperative_verb
Else if (suffix of term in the list of Suf2V)
Then

```

Figure 3. Pseudo code of Arabic Verbs

3.4.2 How to Determinate the Noun type

Nouns are all the proper nouns that are clarified the names of people, animals, and places. Cities...etc. In general, most of the nouns are preceded by (ال, لل, وال, بال, ولل). In addition, numbers, prepositions, adverbs of time, adverbs of place, and calling tools, may precede the noun sometimes. However, some nouns are followed by suffix too.

```

If (prefix of term In the list of PreN|| suffix of term in the list of SufN)
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

A list includes various Jordanian slang words that are not existed in the classical language. In this list, we have all the synonyms of each classical word. Therefore, we use it mainly in the mapping process.

3.4.4 Non- Arabic List

The Jordanian slang contains many non -Arabic words that are considered as results of many factors like;

colonialism, geographical nature of the city, and the political transactions. One of the languages that affected the slang Arabic language is Turkish one or the Ottoman Empire. For example; the word “درابزين” means ladder. Not only Turkish colonization, but also the foreign one affected too, for instance; the word “طريزه” means table. In addition, Persian also must take a big percentage while talking about such an issue especially in the Levant countries like the word “بيجاما” which means Pajamas; or the sleeping dress. Therefore, this algorithm compares the existed words in queries with the non -Arabic words.

3.5 Validation

This section is part of the system that includes the information restoring of the slang language. It aims to check each input query; first by checking the grammar that were previously entered into the system. Then, the structure for both the input query is compared with the existed structure inside the grammar. If its arrangement is right, the system goes directly to the next step, but if it is wrong, it gives the user an order to write a new query. Table (3) shows 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 (V1P)	* _
Present Verb (V2P)	*?
Imperative Verb (V3I)	*!
Noun (W)	N
Question (ASK)	\$
Numbers, Jar, Calling, Zarf (WPreN)	&
Mawsool, Jazem, Naseb (WPreV)	@
Sign Name (SN)	#
EN w Akhwatha (EN)	?
KAN w Akhwatha (KAN)	P
Pronoun (Pro)	!
Five Name (FN)	_
Negative (Neg)	%
Atf	*0%

The explain of the validation:

Table 4. The Format of Arabic Statement

Example 1	words	مين	كتب	الشعر
	symbol	\$	* _	N
Example 2	words	الشعر	مين	كتب
	symbol	N	\$	* _

If the user input “مين كتب الشعر”, after making a comparison between this structure and the known formula in the grammar (ASG), the system looks to see if this format is right or not. The system accepted this query, but after checking another query (الشعر مين كتب), this assigned structure does not match any kind of the structures. Thus, it will be automatically refused, and the user will be directly asked to write a new query in another way.

3.6 Mapping Slang to Classical

This step is considered as one of the Slang Arabic Information Restoring System. Its main goal is to convert words from slang into classical. The process of mapping for all the query words is done by mapping verbs especially the affixes that are related to the slang verbs. They are converted to the equivalent affixes that related to the classical

form. The look up table is very useful where it is used to convert some Arabic Jordanian Slang words. Although this process may be seen as a kind of expensive, but it gives a lot of beneficial clear results.

3.6.1 Verbs Mapping

To convert verbs from slang to classical, we need to make some changes in the prefixes of slang. They are exchanged with the prefix of the classical. In addition, the suffix of the slang is exchanged with the suffix of the classical too.

Examples of Affixes in the Past, Present and Imperative Tense

Table (5 & 6) Show how to convert Slang Present verb to their equivalent classical language, by defined the Slang Prefix and suffix for Present verb and their equivalent classical ones.

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

Present slang prefixes	examples	Present classical prefixes	Examples
بي	بيدرس	ي	يدرس
بت	بتدرس	ي,ت	يدرس,تدرس
بن	بندرس	ن	ندرس
ب	بدرس	ي,ا	يدرس, يدرسا

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

Present slang suffixes	examples	Present classical suffixes	Examples
ي	بنكتبي	ين	نكتبين
وا	بنكتبوا	ون/ان	نكتبون/ نكتبان
ن	بنكتبن	ن	نكتبن

Table (7) shows how to convert Slang Imperative verb to their equivalent classical language, by defined the Slang suffix for Imperative verb and their equivalent classical ones.

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

Imperative slang suffixes	examples	Imperative classical suffixes	examples
ي	ادرسى	ي	ادرسى
وا	ادرسوا	وا, ا	ادرسوا, ادرسا
ن	ادرسن	ن	ادرسن

In Jordanian slang language, the imperative prefix stable like the imperative in classical, it is (ا), and there is no prefix of past tense in Jordanian slang language similar to classical language.

Table (8) shows how to convert Slang Past verb to their equivalent classical language, by defined the Slang suffix for Past verb and their equivalent classical ones.

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

Past slang suffix	examples	Past classical suffixes	examples
ت	لمست	ت	لمست

ن	لمسن	ن	لمسن
تن	لمستن	تن	لمستن
وا	لمستوا	وا, ا	لمستوا, لمستا
تي	لمستي	ت	لمست
نا	لمسنا	نا	لمسنا

3.6.1 Nouns Mapping

Most nouns are unchangeable; constant where the process of converting is not really needed. However, there are few of Jordanian slang words that should be converted by taking help from the look up table. Some foreign words really need to look at the non-Arabic list. Both the affixes of the slang and the classical ones are very similar.

Example explains How the System Works

- 1- I/q: "اليش بيتحرك الحبل لحاله".
- 2- Determination type of terms:
- 3-

Table 9. Each Terms Type

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

1- Check the validation:

In this case its true format \longrightarrow 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 (لحاله) in slang has the synonym meaning in the classical which is (لوحده), and we have the question term (اليش؟) that has the same meaning of (لماذا؟).

b- Then, we check the remain terms in non-Arabic list, and we do not find any word.

c- Now, correct the entire query; لماذا بيتحرك الحبل لوحده.

Here we have a verb (بيتحرك) that has a prefix (بي), we will replace it with present prefix classical form (ي). Then it is converted from (بيتحرك) to become (يتحرك), and (حبل) still the same without changing (حبل). Now we find that we got the classical query:

لماذا يتحرك الحبل لوحده؟

d- Remove the stop words like (لماذا).

e- Now, search in the classical document.

- Second way; search in the slang documents:

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

b- Now search in the slang documents about (بيتحرك الحبل لحاله).

3- Now, show all the results in both the classical and the slang.

4. Results

By using C# language, we built this system under the Microsoft visual studio of 2013. The content of searching or the dataset includes a list of stored documents in Arabic language both slang and standard. Two main stores are used in the process of building the SQL of Microsoft server 2012. The first includes about 568 of documents that are written in Arabic slang and the main concepts in it are spoken which are taken from the slang one and from the forums too. The second includes almost 493 classical documents that are taken from the social data or from articles. Two stop words also are used with 445 classical ones. The list is collecting from a program of a marathon, but the

second is from a group of a hundred collected slang stop words.

It is very important to indicate that, we collected many documents for both slang and classical mainly from the social media; (Facebook, twitter, Instagram, etc.). The collected document was made manually.

4.1 Measuring Performance

In any new system, it is very significant to put light on performance by comparing it with other previous ones. However, the most difficult thing in this system is the lack of slang documents that is used because it is relatively new.

In the process of retrieval, the instances are the documents and the task is to return a group of relevant documents that are given a special search concepts. In other words, I mean to assign to each document one of two categories either relevant or irrelevant. In the case of the relevant documents, they are the documents that belong to the relevant category while the recall is defined as the group of relevant documents that are retrieved by a search they are divided by the total number of existing documents.

Here are some main measure performances that are used basically in the information retrieval system:

A. Precision

It is defined as the number of documents that very relevant to the users query in the whole retrieved documents. Its value is in a range from 0.1 until 1.0. Thus, when it is 1.0 this means that relevant documents (Nwesri and et al., 2008; Kowalski and et al., 2000).

Precision =	$\frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$
--------------------	---

B. Recall

The relevant retrieved document is very relevant to the documents in the database. Its value is recorded between ranges of 0.1 to 1.0. In addition, when its value is 1.0, the relevant documents expected to be (Nwesri and et al., 2008; Kowalski and et al., 2000).

Recall =	$\frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$
-----------------	---

C. F-Measure

It is the mean of harmony between both precision and recall, where its values are 0.0 and 1.0, when it is 0.0 we can expect its lowest values, but when it is 1.0, this means it has reached its highest results out of (Nwesri and et al., 2008; Kowalski and et al., 2000).

4.2 The Experimental Results

The efficiency of this new system can be done in similar behavior of both:

1. By comparing both results of slang and classical queries to apply it on classical, slang, and the hybrid dataset's content.
2. By comparing the values of precision in slang queries with the value of precision first twenty retrieving documents by applying it on Google and yahoo.

5. Decision of Results

5.1 Testing Using Jordanian Slang Language- based over Jordanian Slang Dataset

The system has been tested using Slang – 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 also shown.

Table 10. Precision, Recall and F-measure values after testing Jordanian slang dataset using Jordanian Slang-based Arabic queries

Number	The Query	Precision	Recall	F- measure
1	شو هي اضرار الدخان ؟	0.67	0.65	0.66

2	قديش عدد محافظات الاردن ؟	0.62	0.58	0.60
3	شلون بنحل عجقة السير ؟	0.70	0.56	0.62
4	شلون انزل وزني؟	0.79	0.74	0.76
5	شو احسن جامعه بالاردن ؟	0.66	0.56	0.61
6	شو اسباب البطالة ؟	0.69	0.59	0.64
7	شو اسباب الطلاق ؟	0.68	0.68	0.68
8	شو اسباب الكحة ؟	0.63	0.69	0.66
9	ادرسوا عشان تنجحو .	0.65	0.76	0.70
10	بدي احلى تصميم لفصل بين .	0.63	0.66	0.64
11	مين اللي اخترع الكهربية ؟	0.71	0.71	0.71
12	شو التويتر ؟	0.75	0.77	0.76
13	ليش التلفون بيطفي ؟	0.56	0.54	0.55
14	ليش السيارة بتشتغلش ؟	0.58	0.55	0.56
15	الاغراض اللي بيحتاجها الاولاد .	0.70	0.78	0.74
16	انواع الفواكه والخضار .	0.55	0.67	0.60
17	ياريت الزمان يرجع .	0.45	0.40	0.42
18	نكت بتضحك .	0.54	0.68	0.50
19	صلطة الفواكه .	0.78	0.62	0.69
20	الصرصور من الحشرات ؟	0.82	0.43	0.56
21	هدية لزواج .	0.43	0.50	0.46
22	الحيوانات اللي بتعيش بالبيت .	0.84	0.55	0.66
23	حلويات مشهورة بالاردن .	0.77	0.62	0.69
24	فواكه مشهوره بالاردن .	0.78	0.75	0.76
25	تركيا من اجمل المنافق السياحية .	0.55	0.72	0.62
26	ارتفاع اسعار اللحوم بالاردن .	0.40	0.64	0.49
27	اكثر الشوارع عجقة بالاردن .	0.57	0.69	0.62
28	اكثر مرض منتشر عند الخنايره .	0.76	0.56	0.64
29	شو الحل انسرق حسابي ؟	0.52	0.64	0.57
30	وين بتطلع الخبيزة ؟	0.49	0.20	0.28
31	مين اول زلمه بالعالم ؟	0.28	0.60	0.38
32	شو استخدامات التلفون ؟	0.80	0.69	0.74
33	قديش عدد الكواكب ؟	0.84	0.53	0.65
34	ايتمه بتبلش عطلة الربيع ؟	0.60	0.60	0.60
35	جيف بتعملي الكيك ؟	0.74	0.53	0.62

36	وين بتعيش الارانب ؟	0.51	0.55	0.53
37	شلون اعمل بحث ؟	0.73	0.75	0.74
38	شلون بطبخ سمك ؟	0.48	0.70	0.57
39	شلون بنزل فيديوهات ؟	0.45	0.65	0.53
40	ليش بنشرب العصير ؟	0.53	0.63	0.58
41	شو اخر فيلم بالسنيما ؟	0.69	0.76	0.72
42	الاعراض اللي بتحتاجها النسوان ؟	0.66	0.65	0.65
43	اليس عشان ما تبرد ؟	0.61	0.43	0.50
44	مين اكثر ناس بتارجل ؟	0.45	0.66	0.54
45	شو يدك تدرس بالجامعه ؟	0.43	0.71	0.54
46	قديش عدد القارات ؟	0.62	0.61	0.61
47	اكثر دوله بتزرع زيتون .	0.59	0.70	0.64
48	جيف بتخطي فسطان ؟	0.60	0.60	0.60
49	باي قاره الاردن ؟	0.75	0.64	0.70
50	مين بنات الرسول ؟	0.84	0.63	0.72

5.2 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 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 Arabic classical Language-based Arabic queries

Number	The Query	Precision	Recall	F- measure
1	ما هي اضرار الدخان ؟	0.75	0.58	0.65
2	كم عدد محافظات الاردن ؟	0.65	0.67	0.66
3	كيف نحل ازمة السير ؟	0.70	0.66	0.68
4	كيف انزل وزني؟	0.72	0.72	0.72
5	ما هي افضل جامعه بالاردن ؟	0.66	0.73	0.69
6	ما اسباب البطالة ؟	0.74	0.68	0.71
7	ما اسباب الطلاق ؟	0.64	0.77	0.70
8	ما اسباب الكحة ؟	0.59	0.71	0.64
9	ادرسوا كي تتجحوا .	0.67	0.42	0.52
10	اريد اجمل تصميم فساتين .	0.53	0.69	0.60
11	من الذي اخترع الكهرباء ؟	0.64	0.55	0.59
12	ما هو التويتتر ؟	0.76	0.65	0.70
13	لماذا التلفون يغلق ؟	0.49	0.75	0.59

14	لماذا السيارة لا تشتغل ؟	0.57	0.58	0.57
15	الاعراض التي يحتاجها الاولاد .	0.61	0.69	.650
16	انواع الفواكه والخضار .	0.63	0.64	.630
17	ياليث الزمان يعود .	0.39	0.40	0.39
18	نكت تضحك .	0.62	0.63	0.62
19	سلطة الفواكه .	0.80	0.57	0.67
20	الصرصور من الحشرات ؟	0.82	0.43	0.56
21	هدية لزواج .	0.35	0.49	0.41
22	الحيوانات التي تعيش في البيت .	0.83	0.53	0.65
23	حلويات مشهورة بالأردن .	0.79	0.62	0.69
24	فاكهة مشهورة بالأردن .	0.69	0.70	0.69
25	تركيا من اجمل المنطق السياحية .	0.52	0.60	0.56
26	ارتفاع اسعار اللحوم بالأردن.	0.31	0.63	0.42
27	اكثر الشوارع ازمة بالأردن.	0.71	0.56	0.63
28	اكثر مرض منتشر عند كبار السن .	0.62	0.52	0.57
29	ما الحل انسرق حسابي ؟	0.50	0.70	0.58
30	اين يطلع الخبيزة ؟	0.54	0.25	0.34
31	من اول رجل بالعالم ؟	0.33	0.62	0.43
32	ما استخدامات التلفون ؟	0.78	0.73	0.75
33	كم عدد الكواكب ؟	0.84	0.58	0.69
34	متى تبدأ عطلة الربيع ؟	0.80	0.55	0.65
35	كيف تعمل الكيك ؟	0.77	0.51	0.61
36	اين تعيش الارانب ؟	0.55	0.50	0.52
37	كيف اعمل بحث ؟	0.73	0.63	0.68
38	كيف بطبخ سمك ؟	0.56	0.74	0.64
39	كيف بنزل فيديوهات ؟	0.51	0.65	0.57
40	لماذا نشرب العصير ؟	0.45	0.58	0.51
41	ما اخر فيلم بالسنيما ؟	0.68	0.68	0.68
42	الاعراض اللي تحتاجها النساء ؟	0.70	0.60	0.65
43	البس كي لا تبرد ؟	0.55	0.36	0.44
44	من اكثر ناس بتأرجل ؟	0.56	0.64	0.60
45	ماذا تريد تدرس بالجامعة ؟	0.47	0.72	0.57
46	كم عدد القارات ؟	0.62	0.54	0.42
47	اكثر دوله تزرع زيتون .	0.59	0.73	0.65

48	كيف تخيطين فستان؟	0.70	0.59	0.64
49	بأي قاره الاردن؟	0.68	0.60	0.64
50	من بنات الرسول؟	0.81	0.54	0.65

Figure (5), reveals the results of the recall precision, and it announces that the rates of the slang are very close to the ones of the classical precision due to my use of special rules of slang that we added to our researching system. In addition, we have put some rules to make the conversion process, which makes inversion from slang to classical. Therefore, the results of precision are recorded about 0.63 for both researches. While the results of the recall in this researching system are also recorded as 0.62 and 0.60 respectively for both slang and classical. They are highly considered as good proportions in retrieving the relevant documents. Thus, they are convincing reasons to give the system a good evaluation.

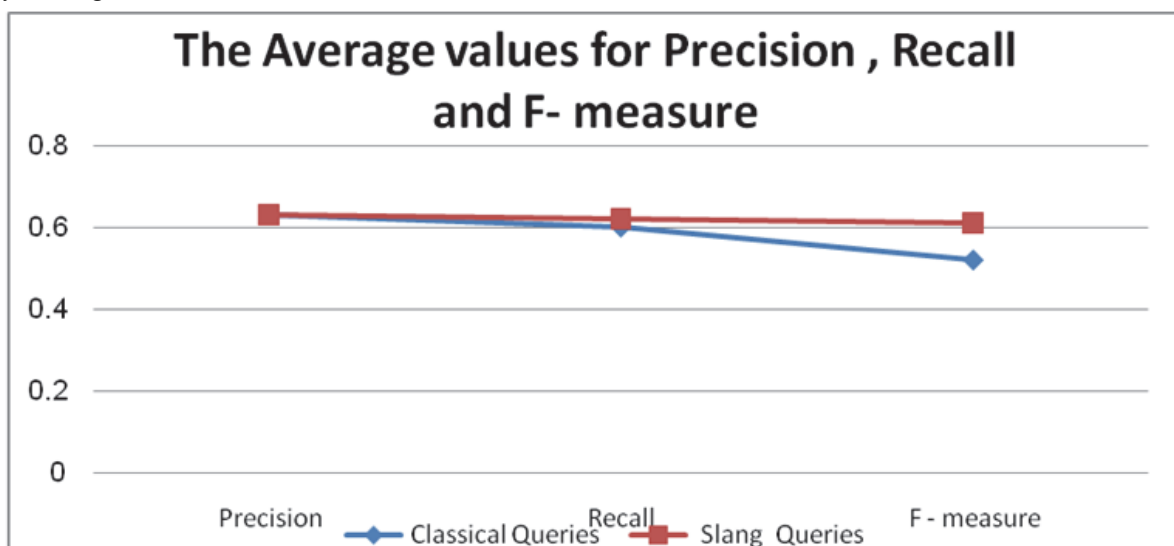


Figure 5. The average values of precision, recall and F- measure values after testing for both slang and classical queries

5.3 Comparing Research System with Google and Yahoo Search Engine

To make it clearer and to prove the research system efficiency, we applied the group of the slang queries, which we mentioned in table (10) on both previous engines. Thus, the first 20 retrieved results were counted in both to make a comparison. This comparison is made between results, and then it is obtained in the Research system through applying the slang queries.

To be honest, which helped this comparison to be well done is that both engines; Yahoo and Google contain documents of both types the slang and the classical.

In Table (12), the final values of precision in both engines are mentioned. Actually, we used precision in this comparison because of its high possibility of being counted; I mean to count its relevant results, but of course after applying the queries. Unlike it, the recall can never be counted because its relevant stored documents in their databases are unknown.

This step is a very important one to us since we worked hard to search about each query in both Google and yahoo, and to find the first 20 results about the entered queries. We figured out whether each one is classified as a relevant one or not. Finally, we divided the relevant document on all the retrieved ones where the result was a value precision 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	0.69	0.40
2	0.62	0.49	0.20
3	0.70	0.42	0.10
4	0.79	0.50	0.20
5	0.66	0.38	0.10
6	0.69	0.75	0.30
7	0.68	0.45	0.20
8	0.63	0.45	0.35
9	0.65	0.50	0.20
10	0.63	0.35	0.10
11	0.71	0.80	0.50
12	0.75	0.65	0.40
13	0.56	0.25	0.10
14	0.58	0.35	0.10
15	0.70	0.45	0.20
16	0.55	0.65	0.25
17	0.45	0.25	0.00
18	0.54	0.70	0.40
19	0.78	0.70	0.50
20	0.82	0.50	0.20
21	0.43	0.00	0.00
22	0.84	0.58	0.30
23	0.77	0.30	0.00
24	0.78	0.40	0.10
25	0.55	0.60	0.20
26	0.40	0.30	0.00
27	0.57	0.50	0.15
28	0.76	0.00	0.00
29	0.52	0.10	0.00
30	0.49	0.55	0.10
31	0.28	0.45	0.00
32	0.80	0.50	0.20
33	0.84	0.80	0.50

34	0.60	0.65	0.15
35	0.74	0.60	0.20
36	0.51	0.70	0.40
37	0.73	0.50	0.10
38	0.48	0.50	0.30
39	0.45	0.40	0.10
40	0.53	0.50	0.30
41	0.69	0.35	0.10
42	0.66	0.25	0.00
43	0.61	0.15	0.00
44	0.45	0.10	0.00
45	0.43	0.25	0.00
46	0.62	0.75	0.40
47	0.59	0.70	0.30
48	0.60	0.65	0.30
49	0.75	0.70	0.30
50	0.84	0.85	0.40

Figure (6) indicates a noticeable comparison in the precision's resulting averages that were shown in Table (12). Eventually, we can see how the Research system shows better values in precision, and they even have higher rates of efficiency than any other searching engines such as; Google and yahoo. In this research, we do not deny that both of them support the process of searching in Arabic, but the results in the Research system are better than theirs, especially while talking about the use of written slang queries.

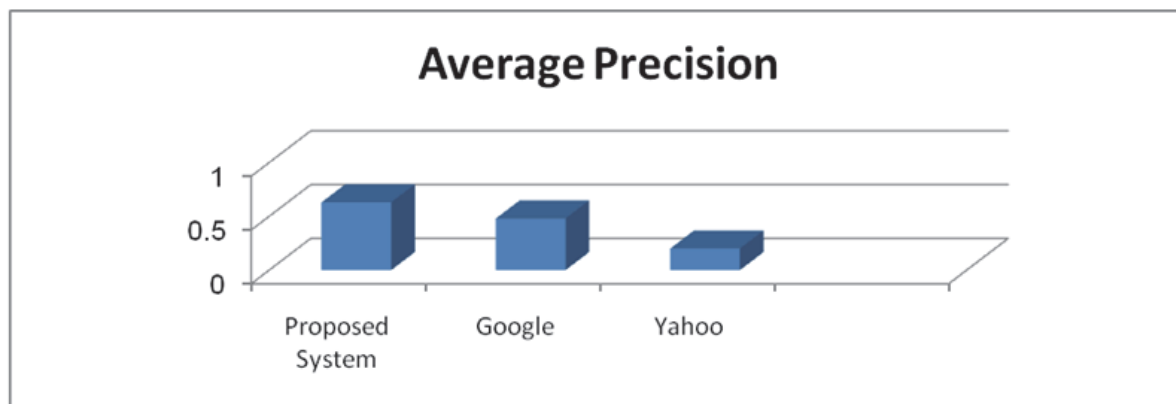


Figure 9. Comparison Results between the average precision values given by Research System, Google and Yahoo Search Engine

6. Conclusion and Future Directions

6.1 Conclusion

This research concerns about Arabic information retrieval process (IR) through providing a new searching system using Jordanian slang queries. By using the Research Arabic slang grammar (ASG), results show a high accuracy

resulted in the retrieved documents while giving hand to the user's query to do its retrieve classical ones.

In addition, Arabic classical documents recorded close rates in three stages of: precision, recalling, and f-measuring, in both slang and classical. In addition, this is something can be easily justified because the Research conversion process makes inversion from the slang concepts to the equivalent terms in classical.

Not only that, but also the slang document is mainly used in searching for both slang and classical queries. In addition, the results also show that using the slang-based queries beat the use of classical ones, but for sure not the opposite. Moreover, results also indicate that the use of slang-based queries is extremely more useful and better than the use of classical ones specifically in the retrieving document process.

It is clearly seen that both queries are used for searching, and both kinds can be used to be entered as inputs. To sum up, results show that using slang queries is better than using the classical one.

6.2 Future Direction

1. Expanding the Research system will be done to combine it to other foreign languages.
2. Categorizing the field of searching will be used to make the process more precise; for instance: politics, geography, etc.
3. Find solution to changing letters in Jordanian slang like "ق" to "أ".
4. Supporting the system by multimedia tools like adding pictures, videos, voices etc.
5. Helping people with disability in the process of searching through making it easier by allowing them to make voice record especially for those who have sight impediments and handy cut, and it will be very useful since spoken slang is better than writing.

References

- Abuata, B., & Al-Omari, A. (2014). A Rule-Based Stemmer for Arabic Gulf Dialect. *Journal of King Saud University– Science*. September. <http://dx.doi.org/10.1016/j.jksuci.2014.04.003>
- Ahmed, F., & Nurnberger, A. (2008). Arabic/English word translation Disambiguation approach based on naïve Bayesian classifier. In: Proceedings of the *international multi conference on computer science and information technology (IMCSIT)*; October 20–22; Wisia. pp. 331–338.
- Al Kharashi, I., & Al Sughaiyer, I. (2004). Performance Evaluation of an Arabic Rule-Based Stemmer. Proceedings of the 17th *national Congress of computer informatics in the service of pilgrims*. 2004 April; King Abdulaziz University, Madinah. 405-414.
- Al-Gaphari, G., & Al-Yadoumi, M. (2010). A method to convert Sana'ani accent to modern standard Arabic. *International Journal of Information Science and Management*, 8(1).
- Almaman, 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. <http://dx.doi.org/10.1109/ICCSPA.2013.6487247>
- Al-Shalabi, R. (2005). Pattern-based stemmer for finding Arabic roots. *Information Technology Journal*, 4(1), 38–43.
- 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., & Al-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*, 41(4). The Author(s) 2015 Reprints and permission: sagepub.co.uk/journalsPermissions.nav. <http://dx.doi.org/10.1177/016555151515610513>
- Benajiba, Y., & Rosso, P. (2007). Arabic question answering [Diploma]. *Technical University of Valencia*. Valencia, Spain, September.
- Carvalho, D., Bessa, M., Magalhães, L., & Carrapatoso, E. (2018). Performance evaluation of different age groups for gestural interaction: a case study with Microsoft Kinect and Leap Motion. *Journal Universal Access in the Information Society*, 17(1), 37-50, Springer-Verlag Berlin, Heidelberg. <http://dx.doi.org/10.1007/s10209-016-0518-4>

- Diab, M., Pasha, A., Al-Badrashiny, M., Altantawy, M., Habash, N., Pooleery, M., Rambow, O., & Ryan, M. (2013). DIRA: Dialectal Arabic Information Retrieval Assistant; *The Companion Volume of the Proceedings of IJCNLP 2013: System Demonstrations, pages 13–16, Nagoya, Japan, 14-18 October 2013*.
- 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. <http://dx.doi.org/10.1109/IIT.2007.4430403>
- 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. ISSN: 2372-479X (Print) 2372-4803 (Online) Copyright © the Author(s). All Rights Reserved. Published by American Research Institute for Policy Development. <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.
- Habash, N., & Rambow, O. (2006). MAGEAD: A Morphological Analyzer and Generator for the Arabic Dialects. <http://dx.doi.org/10.3115/1220175.1220261>.
- 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. <http://dx.doi.org/10.1109/ISPS.2011.5898874>
- Kowalski, J., & Maybury, T. (2000). *Information Storage and Retrieval Systems: Theory and Implementation. 2nd Edition. United States. Spring 2000*. ISBN: 0-7923-9926-9.
- 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*.
- Nwesri, A. (2008). Effective retrieval techniques for Arabic text. PHD Thesis. *School of Computer Science and Information Technology, RMIT University*; May.
- Ogheneovo, E., & Japheth, R. (2016). Application of Vector Space Model to Query Ranking and Information Retrieval, 6(5). *International Journal of Advanced Research in Computer Science and Software Engineering Research Pape*. Retrieved from www.ijarcsse.com
- 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. <http://dx.doi.org/10.1109/ICTTA.2008.4530003>.
- Shatnawi, M., Bani Yassein, M., & Mahafza, R. (2012). A framework for retrieving Arabic documents based on queries written in Arabic slang language *Journal of Information Science*, 38(4), 350 –365. The Author (s) Reprints and permission: sagepub.co.uk/journalsPermissions.nav. <http://dx.doi.org/10.1177/0165551512442480>
- 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. <http://dx.doi.org/10.4038/jula.v17i2.6649>

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/4.0/>).