Volume-4, Issue-3, March-2015 • ISSN No 2277 - 8160

Research Paper | Information Systems

# Computational Method for Accurate Classification of Arabic Texts Based on Arabic Phonetic Transcription

Dr. Ahmad Abdulqader Abuseeni

Department of Management Information Systems, Taibah University, Medina, Saudi Arabia

**ABSTRACT**

*Arabic auto translation is challenging not only because of the inter-language transfer errors that may result, but also because of the lack the absence of an effective, accurate and sufficient retrieval system. In the Arabic language, many words are written in the same letters but have different phonetic transcriptions and meanings. As such, Arabic has a very complicated morphological structure with prefixes, infixes, and suffixes. In this focus of this scientific paper primarily revolves on the analysis of Arabic phonetic transcription based on a Bayesian Method support structure as well as Support Vector machine algorithms to classify Arabic texts. In this analytical paper, analysis focuses on systems than are supported by word pronunciation and diacritical marks. In addition to vowel letters, ablaut and slurring voice units, this paper takes into consideration the diacritical marks which make classification of Arabic texts based on Arabic phonetic transcription more efficient. To this end, experiemtns in this research highlight promising outcomes.*

KEYWORDS : Arabic Text Classification, Computation, Naïve Bayesian Method, Phonetic Transcription Form.

## 1 Background Synopsis and Significance of the Study

### 1.1 Introduction

Odden (2008, p.2) notes that phonology is the theoretical study of how sounds are used in a language to encode meaning by governing the rules of pronunciation. On the other hand, pronunciation can be highlighted as a skill and the act of producing sounds of speech that enables an individual to use spoken words in the correct stress, rhythm, articulation and intonation patterns with reference to a standard of correctness such as Received Pronunciation. This is in contrast to phonetics where Carr (2003) refers to it as the production of the sounds of speech within a language.

Communcation through speech and language is possibly the most important function in the human body. It allows and caters for a variety of commands and understanding through vocal speech patterns. Here, Arabic is one of the world's oldest languages. It is spoken across the Middle East, North Africa and the Horn of Africa. A Central Scemitic language, Arabic is closely related to Aramaic and Hebrew. Arabic is spoken by approximately 420 million speakers and is ranked as the sixth official language of the world (UN, 2013).

There are 10.5 million Arabic speakers with access to the Internet, compared to 287.5 million English speakers (Abusini, 2005). Unfortunately, efforts to improve Arabic information search and retrieval, in comparison to English or French, are limited and modest. The barrier to text processing advancements in Arabic highlights the very complicated morphological structure of the Arabic language. This is perhaps due to the the fact that Arabic word variants are formed by the usage of affixes (prefixes, infixes, and suffixes) and its morphological language suffers from the unavailability of a standard Arabic translation algorithm (Abusini and Abusini, 2010). Other problems arise from the wide range of articles, conjunctions, prepositions, prefixes, and suffixes that could be attached to a single word. Furthermore, Larkeym, Ballesteros and Connell (2002) argue that Arabic suffers from the common problem of irregularity of singular and plural nouns, which is not related to simple affixing. In Arabic, diacritical marks appear either above or below letters and play an essential role in many cases when it comes to distinguish semantically and phonetically between two identical words with the same characters (Momani and Faraj, 2007). Consequently, one of the main reasons that the Arabic language suffers from processing errors is neglecting the diacritical marks.

This study suggests a developmental program for Arabic Phonetic transcription rather than dictation writing. The program will accept Arabic scripts with diacritical marks, as a modern way of avoiding errors in dealing with Arabic texts.

### 1.2 Previous Studies

Sebaweh (1968) highlighgted that interest in letter-to-sound relationship rules is historically documented. Arabic was one of a few languages where pronunciation rules were listed and illustrated more than 12 centuries ago. Divay and Vitale (1997) noted that English and other European languages received more attention; where the phonological system was considered of primary importance in business, commerce and education.

Al Saleem (2011) investigated the Naïve Bayesian method (NB) and Support Vector Machine algorithm (SVM) on different Arabic data sets to as a method for accurate classification of Arabic texts. The bases of his comparison are the most popular text evaluation measures. The experimental results against different Arabic text categorization data sets highlighted that SVM algorithm outperforms the NB with regards to all measures. Laila (2006) compared between Manhattan distance and Dice measures using N-gram frequency statistical technique against Arabic data sets collected from several online Arabic newspaper websites. The results showed that N-gram using Dice measure outperformed Manhattan distance.

Further, Alghamdi, Alqhtani and Alqhtani (2003), developed a program which converted Arabic writing symbols to speech sound symbols. The program converts any Arabic text with diacritical markes or without to speech sound symbols, but it has many details and it isn't suitable for language processing studies. Yet, Elshafei, Al-Muhtaseb and Alghamdi (2006) discussed one of the problems facing computer processing of the Arabic text. Oner such dilemma is the absence of the diacritical marks in modern printed text. These researchers argued that native Arabic readers can identify the proper vocalization of the text, but when it comes to computer processing, the computer still needs to be provided with algorithms to mimic the human ability to identify the proper vocalization of the text. Such tool is an essential infrastructure for many applications such as Text-to-Speech and Automatic Translation.

Sawaf, Zaplo and Ney (2001) presented results using statistical methods including maximum entropy to cluster Arabic news articles. The results derived by these methods were promising without morphological analysis. Consequently, at a Geneva 2004 conference, El-Kourdi, Ben Said and Rachidi, illustrated how they applied NB to classify Arabic web data. Their studies concluded an average accuracy rate of 68.78%. Similarly, El-Halees (2006), used Maximum Entropy for TC on Arabic data sets. The results revealed that the average F-measure increased from 68.13% to 80.41% using pre-processing techniques (normalization, stop words removal, and stemming). However, in relation to the F-measure results, the algorithm developed by El-Halees (2007) outperformed other presented text classification algorithms including those presented by El-Halees (2006), El-kourdi et al. (2004) and Sawaf et al. (2001).

Mesleh (2007) used three classification algorithms, namely SVM, KNN and NB, to classify 1445 texts taken from online Arabic newspaper archives. The compiled texts Automated Arabic Text Categorization using SVM and NB 125 were classified into nine classes: Computer, Economics, Education, Engineering, Law, Medicine, Politics, Religion and Sports. Chi Square statistics was used for feature selection. Mesleh (2007) discussed that "compared to other classification methods, their system shows a high classification effectiveness for Arabic data set in terms of F-measure (F=88.11)" (p. 434).

In Hadi, Thabtah and Alhawari (2008) research, NB and KNN were applied to classify Arabic text collected from online Arabic newspapers including Al-Jazeera, Al-Nahar, Al-Hayat, Al-Ahram, and Al-Dostor. The results pointed out that the NB classifier outperformed KNN base on Cosine coefficient with regards to macro F1, macro recall and macro precision measures. Finally, Thabtah, Eljinini, Zamzeer and Hadi (2009) investigated NB algorithm based on Chi Square features selection method. Their experimental results were compared against different Arabic text categorization data sets and provided evidence that feature selection often increases classification accuracy by removing rare terms.

Froud, Lachkar and Ouatik (2013) proposed to evaluate the impact of text summarization using the Latent Semantic Analysis Model on Arabic Documents Clustering in order to solve problems cited above, using five similarity/distance measures: Euclidean Distance, Cosine Similarity, Jaccard Coefficient, Pearson Correlation Coefficient and Averaged Kullback-Leibler Divergence, for two times, witht and without stemming. The experimental results highlighted that the proposed approach effectively addressed the problems of unintelligible information and document length, and thus significantly improved the clustering performance.

### 1.3 Statement of the Problem

As diacritical marks play an important role in many forms of language structure, especially in cases of vowels, ablaut and slurring, unfortuatley, several of the current Arabic classification systems do not recognise and apply these marks. Consequently, many classification errors may occur, especially when the words have the same letters, but the meaning is different. For example, the words "حُب" which means 'love' and "حَب" which means a type of 'grain', are grouped as one in existing classifiers.

### 0.4 Significance of the Study

The study attempts to propose a developmental program for Arabic phonetic transcriptions that would help individuals use Arabic scripts with diacritical marks to convert text into Arabic speech symbols in a phonetic transcription form. Subsequently, this would help indentify and distingusih between words that have the same characters, but with different meanings.

Thus, the importance of this research arises because of the following reasons; (i) Converting Arabic words into their equivalent phonetic transcription forms would help in the study of the Arabic language, including all its details; (ii) it would solve many of the Arabic language problems arising from neglecting diacritical marks and as a result; (iii) converting Arabic text into its basic units (phonetic transcription form) would open the way to improve the current approaches of language processing.

## 2 Phonetic Transcription of Arabic Words

### 2.1 Definition

The smallest units of the Arabic language are letters with diacritic marks. Each letter is written seperately, with its associated diacritic mark. This specific unit is used to measure the internal relationship between the morphological structure of Arabic, while building the Arabic verbs, so the voice units build the morphological units or words (Abusini, 2005).

For example the units of the word " كَتَبَ " are " ك ـَ ت ـَ ب ـَ "

A phonetic transcription form is the structure in which Arabic words are written as they are pronounced, using the smallest unit of Arabic language; letters, while using diacritical marks separately. Any Arabic word can be written in a Phonetic Transcription form, depending on the way each letter is pronounced, because each letter in Arabic language stemmed out of a specific spot in the lip, inside the mouth, or the larynx (Abusini and Abusini, 2010). This is highlighted in Figure 1 and Table 1, respectively.

**Figure 1: Arabic letters and their pronunciation spot in the lip, inside the mouth, and the larynx.**

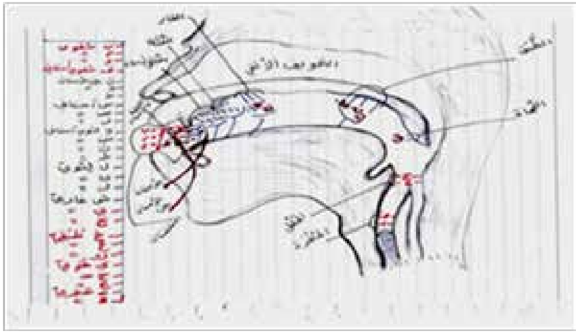


**Table 1: An Example of Arabic words written in a Phonetic Transcription form**

| Word | Phonetic transcription form |
|---|---|
| أكرم | أ ـْ ك ـَ ر ـِ م ـُ |
| إعشوشب | إ ـِ ع ـْ ش ـَ و ـْ ش ـَ ب ـَ |

Therefore, due to word structure and sentence formation in the Arabic language, any Arabic writing form must take into account the diacritical marks (Abusini and Abusini, 2010).

### 2.2 Advantages of Arabic Phonetic Transcription form (PTF) in Dealing with Vowel-ed Words

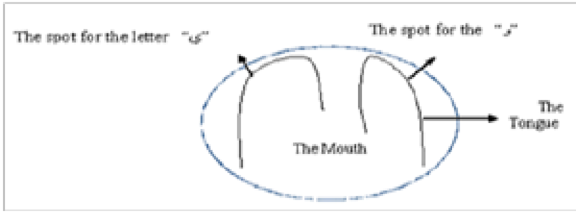
Vowelization: Changing the letters "و" or " ي " to the letter " ا " or vice versa, to facilitate the pronunciation (Abusini and Abusini, 2010)..

Vowels are: و , ي

Example: بَيَعَ      بَاعَ
دَعَوَ      دَعَا
إِسْتَقْوَمَ      إِسْتَقَامَ

There's a certain difficulty in pronouncing the "و" or "ي" in the first or in the end of the word, if there was a diacritical mark before or after it. [23] For example, when you say "وَعَدَ" or يَعَدَ" ", the spot used to pronounce the letter "و" or the letter "ي" will be shown in figure 2.

**Figure 2: Arabic Vowels letters and their pronunciation spot.**



So the letter "و" or "ي» will be changed to " ا" to facilitate the pronunciation, " دَعَوَ " or " دَعَيَ " will become " دَعَا ". By changing the construction of the word because of the Vowels Voice. Nevertheless, using Phonetic Transcription is more suitable than dictation writing, as it is not efficient in dealing with vowels, ablaut, and slurring; voice unit is more accurate and precise (Abusini, 2005). Using basic units of Arabic words may simplify the understanding and the relation between words, and can easily justify the reasons of vowels, ablaut, and slurring.

*Example:*

*The word " [illegible] " converted to the word " [illegible] "*

This can be explained by the following:

The word "وَعَدَ "is written in the phonetic transcription form " و َ ع َ د َ" (note that the letter "و" came between two diacritical marks " ع َ د َ و َ" which makes it difficult to pronounce), so the letter "و" has been eliminated and the word became " ُ ع َ د " which is an equivalent of "اعد ". Therefore, by using the Phonetic Transcription form the reason for converting " دَعْوَ " to " اعد" became easy to understand.

Similarly, phonetic transcription form and the vowels words' voice unit is more accuracy and exactness:

*Example:*

ثقل : ق- ل- ق- ث

The existence of " َ ً " indicate a vowel word.

The word بَاعَ was بَيَعَ and because of the vowel voice the letter "و" has changed to the letter " ا "

*Example:*

عَابَ : ـَ ع َ- َ- ب

The existence of " َ ً " indicate to a vowel word.

The word عَابَ was عَيَبَ and because of the vowels voice the letter " ي" has changed to the letter " ا ".

## 3 The Algorithm

A proposed algorithm will be applied in two phases; (i) the first is the phonetic transcription form in which convert each input word or text to its equivalent phonetic transcription form and (ii) the second phase addrsses the approach in which the classification will be extracted.

### 3.1 First: Phonetic Transcription form algorithm

First: Input the text

Second: Read the input word

Third: Eliminate the prefixes "ال"

Forth: Convert each letter to its equivalent Phonetic Transcription form in based on the following rules:

1) If the letter has " َ " then it is written after it " َ – " in the phonetic transcription form.

2) If the letter has " ِ " under it then it is written after it " ِ " in the phonetic transcription form.

3) If the letter has " ُ "write after it "- ُ " in the phonetic transcription form.

4) If the letter has " ْ " no diacritical mark are written after it.

5) If the letter has " ّ ", then duplicate the letter in the phonetic transcription form.

6) The " أ "and " ء" are to be dealt with as same as letters and written as " ء" in phonetic transcription form.

7) The letter "ا " is represented with " َ ".

8) The letter "ي" is represented with :-ِ- if it does not have diacritical mark on it; as does the previous letter.

9) The letter "و" is represented with : - ُ- ُ does not have diacritical mark on it; as does the previous letter.

10) If the letter does not have any diacritical mark, then write it as it is.

### 3.2 Second: Phonetic Transcription Form (PTF) Categorization:

Phonetic transcription can used to enhance the existing classification method, which eliminate the diacritical marks that plays an important role in many cases. But in this approach, diacritical marks are converted with the world as it is to play the same important role in the phonetic transcription forms.

For example, the word " أَسْوَد " (black) which is the 'color category' and the word "أُسُود" (lions) which is in the 'animals category', are both unfortunately registered in the same classification presses with both words gives returning the same result because the diacritical marks are eliminated.

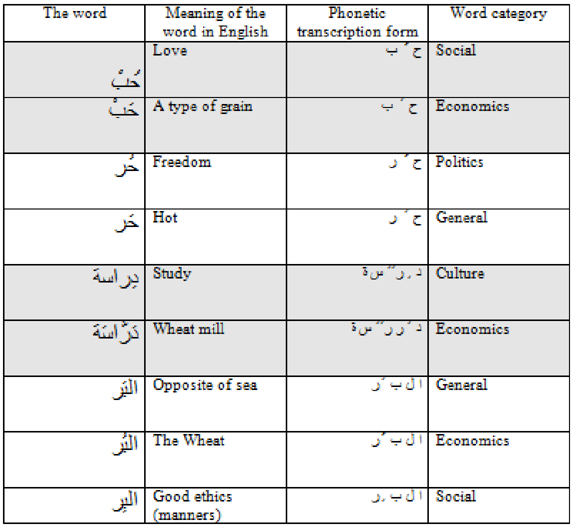
However, using phonetic transcription forms in this program, registry and classification issues will not be realised, for example;

The word [illegible] written as : " أ س و َ د "

The word. [illegible] will be written " أ ُ س ُ ُ د "

Consequently, the program would distinguish between the two similar words and thus rendering the system as more accurate and effective. This is particularly important in the Arabic language, especially when many similar cases can be realised, as highlighted in Table 2, below.

**Table 2: Examples of Different Arabic Words with Different Categories**



| The word | Meaning of the word in English | Phonetic transcription form | Word category |
|---|---|---|---|
| حُبّ | Love | ح ُ ب | Social |
| حَبّ | A type of grain | ح َ ب | Economics |
| حُر | Freedom | ح ُ ر | Politics |
| حَر | Hot | ح َ ر | General |
| دِراسة | Study | د ِ ر ً س ة | Culture |
| دَرّاسَة | Wheat mill | د َ ر ر ً س ة | Economics |
| البَر | Opposite of sea | ا ل ب َ ر | General |
| البُر | The Wheat | ا ل ب ُ ر | Economics |
| البِر | Good ethics (manners) | ا ل ب ِ ر | Social |

After analyzing Table 2, one may easily discover the importance of the diacritical marks to distinguish between the similar words

## 4 Mathematical Approach

NB was applied on a selected Arabic text two times, the first was without using a phonetic transcription form (PTF), and the second was using a phonetic transcription form (PTF ). The following sections, ***Naïve Bayesian (NB)*** will be summarized.

### 4.1 Naïve Bayesian(NB)

The NB is a simple probabilistic classifier based on applying Baye's theorem, and its powerful, easy and language independent method. When the NB classifier is applied to the TC problem we use equation:

P (class document) =

p (class). p (document | class)

P (document)

Where: P (class | document): is the probability of class given a document, or the probability that a given document D belongs to a given class C.

P (document): The probability of a document, we can notice that p (document) is a constance divider to every calculation, so we can ignore it.

P (class): The probability of a class (or category), we can compute it from the number of documents in the category divided by documents number in all categories.

P (document | class) represents the probability of document given class, and documents can be modeled as sets of words, thus the p (document | class) can be written like:

$P \text{ (document} \mid \text{class)} = \Pi i \;\; p \text{ (word } i \mid \text{class)}$

So:

$P \text{ (class} \mid \text{document)} = p \text{ (class) } \Pi i \;\; p \text{ (word } i \mid \text{class)}$

Where: P (word $_i$ | class): The probability that the i-th word of a given document occurs in a document from class C, and this can be computed as follows:

$P \text{ (a word}_i \mid \text{class)} = (Tct + \lambda) / (Nc + \lambda V)$ (4)

Where:

Tct: The number of times the word occurs in that category C

Nc: The number of words in category C

V: The size of the vocabulary table

λ: The positive constant, usually 1, or 0.5 to avoid zero probability [31].

### 4.2 Mathematical Calculation:

I have implemented two manual mathematical calculation experiments on a selected diacritical text from the 'economics category' text and a sample of the 'culture category'. The economics category text was using Naïve Bayesian (N B) on the selected text, and the sample of from the culture category was using voice writing Naïve Bayesian (N B (VF) ) on the same selected text:

#### 4.2.1 NB Experiment:

The input text was the following economics category text:

"ان زراعة البُر من الاهمية بِمَكان  بِحيث يُعتبر
اكبر مُقومات الاقتصاد للدَولة خاصَة عندما يتم
استخدام الدَرَّاسة لحصد البُر ومن ثم جمعه
وتغليفه. "

And the sample of culture category was:

"تُنظم جامِعة المَلِك خالد بالتَعاون مَع المُنظّمة
العَرَبية للتَنمية الاجتماعية في العَاشِر من شَهر
شَوال بِمَدينة اَبْها نِدوَة بِعِنوان  البِر بالوالدين دِراسة
وبحث  تَستَمِر ثَلاثَةُ أيام تهدف النَدوة الى التركيز
على الدراسة وتوضيح اثر البِر بالوالدين في تحسين
ظُروف المجتمع والتقليل من الجرائم ويشارك في
الندوة عدد من التَربَويين والاخصائيين الاجتماعيين
"

First, NB was applied by following the equation:

$$P \text{ (culture} \mid \text{input text)} = p \text{ (culture)} \times \prod_i p \text{ (word}_i \mid \text{culture)}$$

Where: P (word $_i$ | culture): The probability that the i-th word of the input text occurs in a text from culture, and this can be computed as follows:

$$P \text{ (word}_i \mid \text{culture)} = (Tct + \lambda) / (Nc + \lambda V)$$

Where:
Tct: The number of times the word occurs in the culture category.
Nc: The number of words in the culture category.

V: The size of the vocabulary table.
λ: The positive constant, usually 1, or 0.5 to avoid zero probability.

We assume $\lambda = 1$
P (culture) = the number of documents in the culture category divided by documents number in all categories,and this can be computed from table 5.2 as following :

P (culture) = 258 / 1526 = 0.1691. [31]
P (culture | input document) = **P**

$$\text{(culture} \mid \text{input text)} = p \text{ (culture)} \times \prod_i p \text{ (word}_i \mid \text{culture)}$$

$$\mathbf{= 0.1691 \times (1/51 \times 1/51 \times 3/51 \times 3/51 \times 1/51 \times 1/51 \times 1/51 \times 1/51 \times 1/51 \times 1/51 \times 1/51 \times 1/51 \times 1/51 \times 1/51 \times 1/51 \times 1/51 \times 3/51 \times 1/51 \times 3/51 \times 3/51 \times 1/51 \times 1/51 \times 1/51)}$$

**= 0.0175**

So NB decided that the probability that the input text is categorized as culture class = 0.0175.

This result is explained as:
The first 1/51 relates to the number of times the first word in the input text " ان " occurred in the culture sample + λ = (0+1=1) divided by the number of the sample culture words =(51).

The first 3/51 refers to the number of times the third word in the input text " البُر " occurred in the culture sample + λ = (2+1=3), divided by the number of the sample culture words =(51).
Note, that in this case the word " البُر" matched with the word "البر "
And the word " الدَرَّاسة" matched with the word "الدراسة ".

## 4.2.2 Phonetic Transcription Form (PTF) Experiment:

The input text was:

"ان زراعة البُر من الاهمية بِمَكان بِحيث يُعتبر اكبر مُقومات الاقتصاد للدَولة خاصَة عندما يتم استخدام الدَرَّاسة لحصد البُر ومن ثم جمعه وتغليفه. "

The input text was converted to a phonetic transcription form:

"ءن /زرَ َع ة /ب ُر /من /ءهم ِ ِة /
ب ِم َك َ ن /ب ِح ِ ِث /ي ُعتبر /ء
كبر /م ُق ُ ُم َ َت /ءقتص َ َد /لل
دَولة /خ َ َص َة /عندم َ َ /يتم /ء
ستخد َ َم /دَرر َ َس ة /لحصد /
ب ُر /ومن /ثم /جمعه /وتغل ِ ِ
فه /"

And the sample of culture category was:

"تُنظم جامِعة المَلِك خالد بالتَعاون مَع المُنظّمة العَرَبية للتَنمية الاجتماعية في العَاشِر من شَهر شَوال بِمَدينة اَبْها نِدوَة بِعِنوان البِر بالوالدين دِراسة وبحث تَستَمِر ثَلاثَةُ أيام تهدف النَدوة الى التركيز على الدراسة وتوضيح اثر البِر بالوالدين في تحسين ظُروف المجتمع والتقليل من الجرائم ويشارك في الندوة عدد من التَربَويين والاخصائيين الاجتماعيين."

The sample form from the culture category was converted to the phonetic transcription form:

"ت ُنظم /ج َ َم ِع ة /م َل ِك /خ َ َل
د /ب َ َلت َع َ َون /م َع /م ُنظظمة
/ع َر َب ِ ِة /للت َنم ِ ِة /ءجتم َ َع ِ ِ
ة /ف ِ ِ /ع َ َش ِر /من /ش َهر /ش َو َ َ
ل /ب ِم َد ِ ِنة /ء َبه َ َ /ن ِد ُ ُ َة /
ب ِع ِن ُ ُ َ َن /ب ِر /ب َ َل ُ ُ َ َلد ِ ِن /
د ِر َ َس ة /وبحث /ت َست َم ِر /
ث َل َ َث َة ُ /ءي َ َم /
ت َهدف /ن َد ُ ُة /ءلى /تركِ ِز /
على /در َ َسة /وت ُ ُض ِ ِح /ءثر
/ب ِر /ب َ َل ُ ُ َ َلد ِ ِن /ف ِ ِ /تحس ِ ِ
ن /ظ ُر ُ ُف /مجتمع /و َ َلتقل ِ ِ
ل /من /جر َ َىءم /و ِ ِش َ َرك /ف ِ ِ /
ند ُ ُة /عدد /من /ت َرب َو ِ ِ ِن /و َ َ
ل َ َخص َ َىء ِ ِن /ءجتم َ َع ِ ِن /"

P (culture | input document) = **p**

**(culture) × $\prod_i$ p (word $_i$ | culture)**

**=**

**0.1691 × ( 1/51 × 1/51 ×1/51 ×3/51 ×1/51 ×1/51 ×1/51 ×1/51 ×1/51 ×1/51 ×1/51 ×1/51 ×1/51 ×1/51 ×1/51 ×1/51 ×1/51 ×1/51 ×1/51 ×3/51 ×1/51 ×1/51 ×1/51)**

**= 0.0081**

### 4.2.3 Analyzing the Results:

P (culture | input document) **using NB = 0.0175**

P (culture | input document) **using (PTF) = 0.0081**

So NB will be closer in deciding that the input text is a culture text. This was realized because:
In the first experiment (NB) the words " البُر" and " الدَرَّاسة" were matched as the words " البر" and " الدراسة" respectively as the algorithm ignored the diacritical marks.

The input text was:

" ان زراعة البُر من الاهمية بِمَكان بِحيث يُعتبر اكبر مُقومات الاقتصاد للدَولة خاصَة عندما يتم استخدام الدَرَّاسة لحصد البُر ومن ثم جمعه وتغليفه".

The sample of the culture category was:

" تُنظم جامِعة المَلِك خالد بالتَعاون مَع المُنظّمة العَرَبية للتَنمية الاجتماعية في العَاشِر من شَهر شَوال بِمَدينة اَبْها نِدوَة بِعِنوان البر بالوالدين دِراسة وبحث تَستَمِر ثَلاثَةُ أيام
تَهدف النَدوة الى التركيز على الدراسة وتوضيح اثر البر بالوالدين في تحسين ظُروف المجتمع والتقليل من الجرائم ويشارك في الندوة عدد من التَربَويين والاخصائيين الاجتماعيين."

So the algorithm has been matched the word "البُر " with "البِر " and the word " الدَرَّاسة" with " دِراسة" . Addressing this through a mathematical calculation:

p (word " البُر "| culture) ( number of times the word in the input text " البُر " الدَرّاسة" accurse in culture sample + λ divided by the number of the sample culture words).

= (2+1 ) / (51) = **3 / 51**

But in the second experiment (NB[VF]), the words " البُر" and " الدَرَّاسة" were writeen as voice forms " ا ل بُ ر" and " ا ل د َ ر ر َ َ س ة" and thus did not match with the words in the sample culture text " ا ل ب ر " and "ا ل د ر َ َ س ة" .

The input text was
"ء ن / ز ر َ َ ع ة / ب ُ ر / م ن / ء ه م ِ ِ ة /
ب ِ م َ ك َ َ ن / ب ِ ح ِ ِ ث / ي ُ ع ت ب ر / ء
ك ب ر / م ُ ق ُ ُ م َ َ ت / ء ق ت ص َ َ د / ل ل
د َ و ل ة / خ َ َ ص َ ة / ع ن د م َ َ / ي ت م / ء
س ت خ د َ َ م / د َ ر ر َ َ س ة / ل ح ص د /
ب ُ ر / و م ن / ث م / ج م ع ه / و ت غ ل ِ ِ
ف ه /."

And the sample of the culture category was
"ت ُ ن ظ م / ج َ َ م ِ ع ة / م َ ل ِ ك / خ َ َ ل
د / ب َ َ ل ت َ ع َ َ و ن / م َ ع / م ُ ن ظ ظ م ة
/ ع َ ر َ ب ِ ِ ة / ل ل ت َ ن م ِ ِ ة / ء ج ت م َ َ ع ِ ِ
ة / ف ِ ِ / ع َ َ ش ِ ر / م ن / ش َ ه ر / ش َ و َ َ
ل / ب ِ م َ د ِ ِ ن ة / ء َ ب ه َ َ / ن ِ د ُ ُ َ ة /
ب ِ ع ِ ن ُ ُ َ َ ن / ب ِ ر / ب َ َ ل ُ ُ َ َ ل د ِ ِ ن /
د ِ ر َ َ س ة / و ب ح ث / ت َ س ت َ م ِ ر /
ث َ ل َ َ ث َ ة ُ / ء ي َ َ م /
ت َ ه د ف / ن َ د ُ ُ ة / ء ل ى / ت ر ك ِ ِ ز /
ع ل ى / د ر َ َ س ة / و ت ُ ُ ض ِ ِ ح / ء ث ر
/ ب ِ ر / ب َ َ ل ُ ُ َ َ ل د ِ ِ ن / ف ِ ِ / ت ح س ِ ِ
ن / ظ ُ ر ُ ُ ف / م ج ت م ع / و َ َ ل ت ق ل ِ ِ
ل / م ن / ج ر َ َ ئ م / و ِ ِ ش َ َ ر ك / ف ِ ِ /
ن د ُ ُ ة / ع د د / م ن / ت َ ر ب َ و ِ ِ ِ ن / و َ َ
ل َ َ خ ص َ َ ئ ِ ِ ن / ء ج ت م َ َ ع ِ ِ ِ ن /."

So the algorithm here did not match the word "البُر " with "البِر "
And the word "الدَرَّاسة " with " دِراسة"

through a mathematical calculation:
p (word " البُر "| culture) (number of times the word in the input text " البُر " accurse in culture sample + λ divided by the number of the sample culture words).

= (0+1 ) / (51) = **1 / 51**

p (word " الدَرَّاسة "| culture) (number of times the word in the input text " الدَرَّاسة" accurse in culture sample + λ divided by the number of the sample culture words).

= (0+1 ) / (51) = **1 / 51**

It can be therefore concluded that using a PTF will result in more accuracy when dealing with Arabic diacritical text.

## 5 Future Direction

It is believed that converting Arabic text into its basic units (Phonetic Transcription form) will open the way to improve current approaches to language processing or stemming algorithms.

The problem is that most Arabic texts, especially the computer-based ones, do not have diacritical marks. However, there are programs that convert non-diacritical texts to ones with diacritical marks. Several years ago, Google launched a new service called "Tashkeel info" which converted non-diacritical texts to ones with diacritical marks words. The percentage of correctness in the program was good, but there is still room for improvement when creating algorithims for language translations (Diavy, 1997).

Therefore, in the future, it is expected that all the computer-based documents will include diacritical marks and such developments in the future could solve the non-diacritical text problems.

## 6 Conclusion

Morphological analysis is the first step of most normal language processing applications. In Arabic, a lot of errors occur, especially when dealing with vowels as well as in case of classification, particularly when the words have the same letters, but the meaning is different. Such conflicts occur due to the neglect of diacritical marks.

This scientific paper has proposed a new computational method for accurate classification of Arabic texts based on the Arabic phonetic transcription to help to help users deal with Arabic scripts with diacritical marks. It is envisaged this will have a major role in resolving many Arabic information systems' problems arising from neglecting diacritical marks. It can also help Arabic language learners with reading and understanding the relationship between words and ineffective dictation writing, especially

## REFERENCES

Abusini A.A. and Abusini S.M. (June, 2010). An Approach for extracting Arabic word root based on phonetic transcription forms, An Academic Perspective 139. Paper presented at the 14th International Business Information Management Association Conference, Istanbul- Turkey. | | Abusini, S. (2005). Morphological and phonetic reading in Arabic language structure. Zarka Journal for Research and Studies, 7:1, Rabia 2. | | Alghamdi, M., Alqhtani, S., & Alqhtani, S. (2003). Arabic Writing Symbols. Riyadh: King Saud University. | | Alrajehi, A . (2004). Applied Morphology. Beirut: Dar An Nahdha Al Arabiah. | | Al-Saeedi, M. (1999), Awdah Almasalik ila Alfiyat Ibn Male. Beruit: Dar Ihyaa Al Oloom. | | Al Saleem, S. (2011). Automated Arabic Text Categorization Using SVM and NB. International Arab Journal of e-Technology, 2:2, 124-128. | | Al-Shammari, E. and Lin, J. (2008). A novel Arabic lemmatization algorithm, Proceedings of the second Workshop on Analytics for noisy unstructured text data, (pp. 113-118). New York: ACM. | | Benkhalifa, M., Mouradi, A. and Bouyakhf, H. (2001). Integrating WordNet knowledge to supplement training data in semi-supervised agglomerative hierarchical clustering for text categorization. International Journal of Intelligent Systems, 16:8, 929-947. | | Carr, P. (2003). English Phonetics and Phonology: An Introduction. Oxford: Oxford University Press. | | Daniel L. and Pedro, D. (2005). Naive Bayes Models for Probability Estimation, Paper presented at the 22nd International Conference on Machine Learning, Bonn, Germany. | | Divay, M. and Vitale, A. (1997). Algorithms for Grapheme-Phoneme Translation for English and French. Journal of Computational Lingusitics, 23:4, 495-523. | | | El-Halees A. (2006). Mining Arabic Association Rules for Text Classification, Proceedings of the First International Conference on Mathematical Sciences, 15 -17, Al-Azhar University of Gaza: Palestine. | | El-Halees A. (2007). Arabic Text Classification Using Maximum Entropy. Journal Series of Natural Studies and Engineering, 15:1, 157-167. | | | El-Kourdi, M., Ben Said, A., & Rachidi, T. (2004). Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm. Paper presented at the 20th International Conference on Computational Linguistics, Geneva. | | Froud, H., Lachkar, A. & Ouatik, S. (2013). Arabic text summarization based on latent semantic analysis to enhance arabic documents clustering. Retrieved on 13th June 2013 from http://arxiv.org/ftp/arxiv/papers/1302/1302.1612.pdf | | Guo, G., Wang, h., Bell, d., Bi, y. & Greer, K. (2004). A KNN Model-based Approach and its Application in Text Categorization, Proceedings of the 5th International Conference on Intelligent Text Processing and Computational Linguistic, CICLing, LNCS 2945 (pp.559-570) Springer-Verlag. | | Hadi W., Thabtah F., Alhawari S., Ababneh J. (2008) Naive Bayesian and K-Nearest Neighbour to Categorize Arabic Text Data. Proceedings of the European Simulation and Modelling Conference. Le Havre: France (pp. 196-200). | | Hammo, B., Abu-Salem, H., Lytinen, S., & Evens, M. (2002). QARAB: A Question Answering System to Support the Arabic Language, Workshop on Computational Approaches to Semitic Languages, ACL 2002 (pp. 55-65). Philadelphia, PA. | | Joachims T. (1998). Text Categorization with Support Vector Machines: Learning with Many Relevant Features, Proceedings of the European Conference on Machine Learning (ECML), (pp. 173- 142). Berlin: Germany | | Joachims T. (1999). Transductive Inference for Text Classification using Support Vector Machines. Proceedings of the 16th International Conference on Machine Learning (pp. 200-209), San Francisco: Morgan Kaufmann. | | Laila K. (2006). Arabic Text Classification Using N-Gram Frequency Statistics Comparative Study. DMIN, pp. 78-82. | | Larkey, L.S. Ballesteros, L. & Connell, M. (2002). Improving Stemming for Arabic information retrieval. Light stemming and co-occurrence analysis, ACM, Tampere, Finland, pp. 275-282. | | Mesleh, A. A. (2007). Chi Square Feature Extraction Based Svms Arabic Language Text Categorization System, Journal of Computer Science, 3:6, 430-435. | | Momani, M. and Faraj, J. (2010). A Novel Algorithm to Extract Tri-Literal Arabic Root. Journal of the American Society for Information Science and Technology, 61:3, 583-591. | | Moustafa Elshafei, Husni Al-Muhtaseb, and Mansour Alghamdi. (2006). Machine Generatrion of Arabic Diacritical Marks. Paper presented at the International Conference on Machine Learning; Models, Technologies & Applications (MLMTA'06). | | Odden, D. (2008). Introducing Phonology. Cambridge: Cambridge University Press. | | Sawaf, H. , Zaplo,J. and Ney. H. (2001). Statistical Classification Methods for Arabic News Articles, Arabic Natural Language Processing, Workshop on the ACL, Toulouse, France. | | Sebastiani, F. (1999). A Tutorial on Automated Text Categorization, Proceedings of the ASAI-99, 1st Argentinian Symposium on Artificial Intelligence (pp. 7-35). Argentina. | | Sebaweh, A. 1968) An Arabic Book. Cairo: Dar Alkitab Al-Araby. | | Thabtah F., Eljinini M., Zamzeer M., & Hadi W. (January 2009). Naïve Bayesian based on Chi Square to Categorize Arabic Data. Proceedings of The 11th International Business Information Management, Association Conference (IBIMA) Conference on Innovation and Knowledge Management in Twin Track Economies, (pp. 930-935). Cairo: Egypt. | | Thabtah F., Hadi W., Al-shammare G. VSMs with K-Nearest Neighbor to Categories Arabic Text Data. In the World congress on engineering and computer science 2008. (pp.778-781), 22-44 October 2008. San Francisco, USA, 2008. | | UNESCO. (2012), World Arabic Language Day. Retrieved on 12 February 2014. from http://en.wikipedia.org/wiki/UN_Arabic_Language_Day | | United Nations (2013). UN Official Languages. Retrieved from www.un.org on 20th April 2013. | | Wa'el, M., Eljinini, M., Mohammad, A., and Ghatasheh, M. (June 2010). Performance of NB and SVM Classifiers in Arabic Text Data, Business Transformation through Innovation and Knowledge Management, An Academic Perspective 2593, Paper presented at the 14th International Business Information Management Association Conference, Istanbul: Turkey. |