Improving the Performance of Naïve Bayes Algorithm for Arabic Text Categorization

Akram M. O.Al Mashaykhi
Amman Arab University, CIS
Amman, Jordan
akram.othman@aau.edu.jo

Nibras Jamal Abu Aqoulah
Amman Arab University, CIS
Amman, Jordan
aqoulah@hotmail.com

May H. Riadh
Zarqa University, CS
Zarqa, Jordan
may@zu.edu.jo

Abstract- Text Categorization (classification) is the process of classifying documents into a predefined set of categories based on their content, In this paper four techniques are implemented using Naïve Bayes classifier for Arabic text categorization, these techniques are: (TF only ,TF-IDF, Normalized TF-IDF, and N-Gram with N=2 statistical stemmer with threshold similarity 0.8).

The four techniques are evaluated by two test set. The results shows that the Normalized TF-IDF and N-Gram with N=2 statistical stemmer with threshold similarity 0.8 technique has the best accuracy ,the Analysis of Naïve Bayes classifier algorithm showed at least two Advantages:

first it Work well on numeric and textual data and second its easiness in implementation and computation comparing with other algorithms also the work highlighting at least three Disadvantages: first the Conditional independence assumption is violated by real-world data; second its perform very poorly when features are highly correlated and the last disadvantages it does not consider frequency of word occurrences.

Key words- Text Categorization; Naïve Bayes classifier;Arabic categorization; N-Gram.

I. INTRODUCTION

In recent years a huge amount of information that are available through internet, digital libraries and company intranet based application. Robb [1] Many papers focus in dealing with data that are stored in databases and data

Warehouses which are known as structured data. Therefore, communities shift their attentions to these kinds of information sources and forms which are known as unstructured data; this led to Text mining.Text mining is data mining, which are applying to textual data. Data mining is the task of discovering interesting pattern from large amounts of data, where the data can be stored in databases, data warehouse or other information repositories [2].Text mining is the process of finding interesting or useful patterns in a corpus of unstructured textual information Text mining use many techniques that are used in information retrieval, natural language processing and text summarization.Text mining are different from the information retrieval. In information retrieval, the users are looking for known information, but in text mining the users are looking for unknown information. In recent years, many applications of text mining have grown in many areas such as Analysis of survey data, Spam identification, Surveillance, Call center routing, Alias identification.[3]

A. Problem statement

The major difficulty in document categorization is the huge amount of features (distinct word) in the single document in moderate size documents, it is easy to find tens of features, and another difficulty is the complexity in the Arabic language that must deal with it. Another difficulty that is decided which technique will be most suitable for Arabic language when applied with Naïve Bayes classifier algorithm.

B. Arabic document categorization

As the amount of electronic documents increase and the need to extract information and classifying these documents, document categorization task becomes more and more important as described in[4], document categorization(which also knowing as text categorization or topic spotting) is an attempts to replace and save human effort required in performing manual categorization. It is very important task that can help us to classify the documents and find the exact information as the user request. Documents categorization has been used in many fields such as information management, search engines, digital library systems, classifying e-mails, structure search and/or browsing, topic identification and newsgroups classification.[4]

There are many difficulties because of the nature and the complexity of the Arabic language; diacritics that usually left out in the text which create ambiguity in that text represent some of the vowels. In addition, Arabic language

do not use Capitalization for proper nouns that is necessary in documents categorization [5].Many preprocessing techniques are required to the Arabic documents; to make the Arabic documents applicable for the categorization. Some of these pre-processed techniques are tokenizing, stemming and part of speech.[6]

II. THE NAÏVE BAYES PROBABILISTIC MODEL

Thomas Bayes (1702 – 1761) was an English mathematician and Presbyterian minister, known for having formulated a specific case of the theorem that bears his name: Bayes' theorem, which was published. [7] The probability model for a classifier is a conditional model p(C|F1, F2, ..., Fn) over a dependent class variable C with a small number of outcomes or classes, conditional on several feature variables F1 through Fn. The problem is that if the number of features n is large or when a feature can take on a large number of values, then basing such a model on probability tables is infeasible.

The Bayes′ theorem relates the conditional and marginal probabilities of stochastic events C and F:[8]

$$Pr(C|F) = \frac{Pr(F|C)Pr(C)}{Pr(F)}$$

$$p(C|F_1, \ldots, F_n) = \frac{p(C)p(F_1, \ldots, F_n|C)}{p(F_1, \ldots, F_n)}. \quad (1)$$

Where: Pr(C) is the prior probability of hypothesis C; P(F) is the prior probability of training data F; P(C|F) is the probability of C given F and; P(F|C) is the probability of F given C. Using Bayes′ theorem for several feature variables Fn, we can rewrite this as:

$$p(C, F_1, \ldots, F_n) = p(C)p(F_1|C)p(F_2|C, F_1)p(F_3|C, F_1, F_2)\ldots p(F_n|C, F_1, F_2, \ldots, F_{n-1}) \quad (2)$$

In practice we are only interested in the numerator of that fraction, since the denominator does not depend on C and the values of the features Fi are given, so that the denominator is effectively constant. The numerator is equivalent to the joint Probability model (1) which can be rewritten using repeated applications of the definition of conditional probability as:

$$p(C, F_1, \ldots, F_n) = p(C)p(F_1|C)p(F_2|C, F_1)p(F_3|C, F_1, F_2)\ldots p(F_n|C, F_1, F_2, \ldots, F_{n-1}) \quad (3)$$

This means: assuming that each feature Fi is conditionally independent of every other feature Fj for j 6= i and p(Fi|C, Fj) = p(Fi|C) the model (1) can be expressed as[8]:

$$p(C, F_1, \ldots, F_n) = p(C)p(F_1|C)p(F_2|C), \ldots = p(C)\prod_i p(F_i|C) \quad (4)$$

A. Goal and objective

Briefly the main goal and objective of current work can described as follows: Improve Naïve Bayes classifier algorithm after applying the proposed approach on it, Improve the accuracy of Arabic text after applying the proposed pre-processing approach and Determine which technique will be most suitable for Arabic categorization when applied with Naïve Bayes classifier algorithm.

III. METHODOLOGY

The adopted method has three phases each tackle part of the Arabic text categorization process:

A. *The pre-processing phase*

prepares the document for the classification process; this is similar to many natural language processing (NLP) problems where tags and stop words are removed, Stemming is applied to the remaining text in the document this helps reducing the feature space of the problem.[9] In the next step of classification, the documents are split into training and testing documents, the training documents are used to train the system to recognize different patterns of categories, and the testing documents are used to evaluate the system.

B. *Training Phase*

A training set is a set labeled instances used to learn a model.[9] In this paper, the training phase will be processed as shown in Fig 1 which shows the details of the training phase.

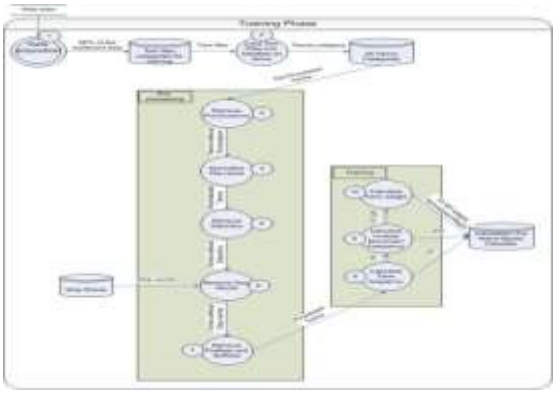

Figure 1 details of the training phase

- Data Acquisition

The data acquisition is the first step, it should be suitable for the objectives of this study, must be in Arabic language. The data is collected manually from many web sites such as (alghad.com), and their format are text files, and organized in five categories (Sport, Art, Health_medicine, Economics and Agriculture),The collected data is divided into 60% as training set and 40 % for the test purpose.

- Load text files into database

In this step all the text files contents will the stored in database, every text file will be loaded in the database as set of terms(words). this process facilitates the following processes.

- Remove Punctuations

Some Arabic documents may contain foreign words, these words are not important for us so these words has to be ignored by the system.[10]so they will be removed because they may be exist in begin or end of terms, (, ' " " . : & @ # % ^ *) (- _ ;).

- Normalize the terms (words)

In this step, all the terms are normalized as the following:1- Normalize آ , إ , أ to ا , 2- Normalize ة to ه

3- Normalize ى to ي , 4- Normalize the sequence يء and the sequence ىء to ئ.

- Remove Diacritics

Some Arabic documents may contain foreign words, these words are not important for us so these words has to be ignored by the system.[10]so they will be removed because they may be exist in begin or end of terms, (, ' " " . : & @ # % ^ *) (- _ ;).

- Normalize the terms (words)

In this step, all the terms are normalized as the following:1- Normalize آ , إ , أ to ا , 2- Normalize ة to ه, 3- Normalize ى to ي , 4- Normalize the sequence يء and the sequence ىء to ئ.

- Remove Diacritics

In this step, removes all diacritics (ً ,ٌ ,ٍ ,َ ,ُ ,ِ ,ّ ,ْ, ~, ٓ), which are marks above or below letters used in orthography and as a sign for the word grammatical case.

- Remove Stop Words

Stop words list is used to identify frequent words that are unlikely to assist in classification such as [غيرك ، هذه انتم ، فانهم] and hence are deleted during pre-processing [13]. The algorithm checks if the word is one of the stop words list that are used here is consists of 1219 word based mainly on Khoja list [6, 9].

- Remove prefixes and suffixes

The algorithm removes a set of prefixes (و , بال , فال , لا , ال , لل). After removing these prefixes it checks if the word length is less than 3 letters. In this case this prefix is considered as a main part of the word and so the removed prefix is returned back to the word.then the suffixes (ه , ي , هما , اما , ات , يه , ته , تي , ان , ون , ين , هم , هن , ها , نا , وا , ام , ان) are recursively removed from the tail of the word. The longest suffix is removed first, then the shorter. This process is recursive because most suffixes are compound of pronouns, gender and number suffixes. For example the word (مكتباتهم) (Their libraries) has a composite suffix (اتهم) which is made from two parts (ات) for feminine plural and the pronoun (هم). Also as done in the previous step the algorithm checks if the word length is greater than 3 letters in order not to remove a main part of the word [6].

- Calculate Term Frequency (TF)

The term frequencyTFt,d of term t in document d is defined as the number of times that t occurs in d.[11]

- Calculate Document Frequency

Rare terms are more informative than frequency terms. In this step, DFt is the document frequency of t the number of documents that contain t, df is a measure of the in formativeness of t.idf , the inverse document frequency of t is defined by:IDF = log (N/dft), The use of log N/DFt instead of N/DFt is to dampen effect of idf .[12]

- Calculate Terms Weight

There are several ways of determining the weight TFt,d of word i in document k, but most of the approaches are based on two empirical observations regarding text[6] : first , the more times a word occurs in a document, the more relevant it is to the topic of the document. Second the more times the word occurs throughout all documents in the collection, the more poorly it discriminates between documents. Thus, the tf-idf weight of a term is the product of its TF weight and its IDF weight Wt,d = (TF x IDF) .[11], And this is the best known weighting scheme in information retrieval. [6], the formula increases with the number of occurrences within a document and increases with the rarity of the term in the collection. [12] Now the training is complete and the system has the ability to classify entire text depending on the information for Naïve Bayes classifier.

C. Test phase

Test set is a data set of unlabelled examples used for testing of the performance of the model learned on the training data set. [12] Fig 2 shows the details of the test phase.

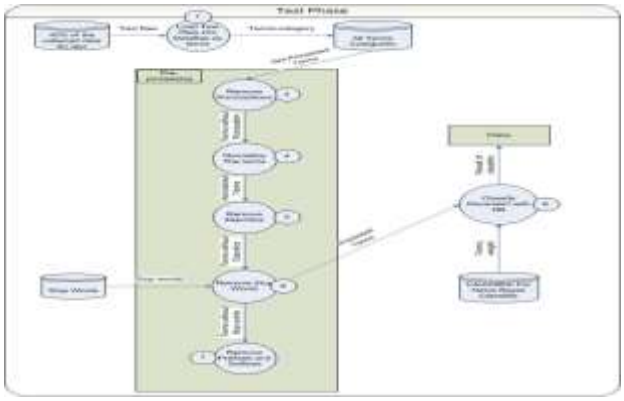

Figure 2 details of the test phase

- Pre-processing for test set

Here some pre-processing is done to the terms such as remove punctuations, remove diacritics, normalize the terms, and finally remove prefix and suffix from the term, these procedures are mentioned and explained before.

- Classify Document

Text Categorization (classification) is the process of classifying documents into a predefined set of categories

based on their content.[4] the techniques that will be used in this paper will be as follows:

- Naïve Bayes Classifier with TF (generative model)

Let D be a document represented as a set of finite terms/roots, D={w1, w2,..., wn}.for each root (term) wk , Compute P(wk/Cj)= (Nk,j +p)/(n+p) (1)

where Nk,j (TF) is the number of times wk occurs in Cj,and n is the number of training documents labeled Cj ,p is the propability of category,Equation (1) make use of the following two assumptions, 1) Assuming that the order of the words does not affect the classification of the document,then :

P(D|Cj)=P({w1, w2,..., wn}|Cj) (2)

Assuming that the occurrence of each word is independent of the occurrence of other words in the document,

P(w1,...,wn|Cj)=P(w1|Cj)*P(w2|Cj)*...*P(wn|Cj) (3)

- Naïve Bayes Classifier with TF-IDF weight

In this technique the NB classifier will use the weight of terms TF-IDF instead of the occurance of terms TF, The modification will be on equation (1) as follow :P(wk/Cj)= (WEIGHTk,j +p)/(n+p) (4)where WIGHTk,j is the TF-IDF weight of wk in Cj, n is the number of training documents labeled Cj and p is the propability of category.Then apply the equation (2) and (3)

- Naïve Bayes Classifier with Normalized TF-IDF weight

The NB classifier will use he Normalized weight of terms TF-IDF instead of TF-IDF weight, the modification will be on equation (5) as follow :P(wk/Cj)= (N_WEIGHTk,j +p)/(n+p) (5)where N_WIGHTk,j is the Normalized TF-IDF weight of wk in Cj. n is the number of training documents labeled Cj ,p is the propability of category.Then apply the equation (2) and (3)

- Naïve Bayes Classifier with Normalized TF-IDF and statistical stemming N-Gram (N=2)

In this technique the NB classifier will use he Maximum normalized weight of similarity terms in the category, the modification will be on equation (6) as follow :P(wk/Cj)= (MAX(N_WEIGHTk,j)+p)/(n+p) (6)where N_WIGHTk,j is the Normalized TF-IDF weight of wk in Cj. n is the number of training documents labeled Cj ,p is the propability of category.Then apply the equation (2) and (3). The similarity is detected using N-Gram algorithm, Similarity measures are determined for all pairs of terms in the Entire Text after Pre-Processing in the test phase,the following example shows how digram similarity between the word (صحي) and the word (صحيا) is measured.1- صح، حي⇒ صحي . (Divided the word into set of digrams each of two adjacent letters)2- Unique digrams ⇒ صح ، حي , 3- صحيا⇒صح ، حي ، يا 4- Unique digrams ⇒ صح ، حي ، يا .

Similarity = 2c/(a+b) =2*2/(2+3)=0.8 m , where a and b are the numbers of unique digrams in the first and the second words. c is the number of unique digrams shared by a and b. The algorithm considers the two words similar if the similarity value is equal or greater than threshold value. Experimental results in [6] show that the most suitable predefined threshold is 0.8.

IV. EXPERIMENT AND EVALUATION

The data is collected manually from many web sites; the data reside in five categories, as said before. In order to test performance of the four techniques, the collection data that contain 489 documents and 76,811 words are divided into training set, and test set, Table 1 shows the number of documents and words for each category in the training set after pre-processing. Table 2 shows the number of documents and words for each category in the test set after pre-processing; Moreover, another data is collected for the purpose of test. This data is collected from different electronic newspapers. Table 3 shows the number of documents and words for each category.

Table1: Number of documents and words for each category for training set

Category	No. of documents	No. of words
Art	36	5,186
Agriculture	40	8,795
Economics	40	33,752
Health_medicine	40	10.188
Sport	40	12.256

Table 2: Number of documents and words for each category for the first test set

Category	**No. of documents**	**No. of words**
Art	10	2,691
Agriculture	9	6,532
Economics	10	2,922
Health_medicine	10	5,870
Sport	10	2,845

Table 3: Number of documents and words for each category for the second test set

Category	**No. of documents**	**No. of words**
Art	**10**	**2,691**
Agriculture	**9**	**6,532**
Economics	**10**	**2,922**
Health_medicine	**10**	**5,870**
Sport	**10**	**2,845**

A. *Evaluation measure*

To accomplish the objectives of this paper, the time that each method takes for the classification task will be

computed and accuracy for each method will be computed to measure the performance , confusion Matrix contains information about actual and predicted classification done by a classification system; performance of such system is

	Agriculture	Art	Economics	Health_medicine	Sport	Total
Agriculture	**39**	0	1	0	0	40
Art	0	**13**	22	1	0	36
Economics	0	0	**40**	0	0	40
Health_medicine	0	0	2	**37**	0	39
Sport	0	0	2	0	**38**	40
Total	39	13	67	38	38	**195**

commonly calculated using the data in the matrix, where the Accuracy = Number of documents that correctly classified to the category / Number of all documents in the test set.

B. Evaluation Result for the four techniques

The test set of data is divided in two sets for the evaluation.

- Evaluation Result for the first test set

This section will be divided into four subsections; each one will present the result of one of the four techniques.

1. Naïve Bayes with TF-IDF

After using the test set by technique one, the confusion matrix that is produced are shown in Table 4.

Table 4 : Confusion matrix for TF-IDF technique

	Agricu_lture	Art	Economics	Health_medicine	Sport	**Total**
Agriculture	**39**	0	1	0	0	40
Art	0	**14**	21	1	0	36
Economics	0	0	**40**	0	0	40
Health_medicine	0	0	2	**37**	0	39
Sport	0	0	2	0	**38**	40
Total	39	14	66	38	38	**195**

Overall classification accuracy achieved over all categories can be calculated as (39+14+40+37+38)/195= 0.861 %.

2. Naïve Bayes Only

After using the test set by technique two the confusion matrix that is produced is shown in Table 5, overall classification accuracy achieved over all categories is (39+13+40+37+38)/195= 0.856 %.

Table 5 : Confusion matrix for Naive Bayes Only technique

3. Naïve Bayes with Normalized TF-IDF

After using the test set in the program by technique two, the confusion matrix that is produced is shown in Table 6.

Table 6 : Confusion matrix for Normalized TF-IDF technique

	Agri cul-ture	Art	Economics	Health_medicine	Sport	Total
Agriculture	**40**	0	0	0	0	40
Art	0	**36**	0	0	0	36
Economics	0	8	**32**	0	0	40
Health_medicine	0	0	0	**39**	0	39
Sport	0	0	0	0	**40**	40
Total	40	44	33	39	40	**195**

The overall classification accuracy achieved over all categories is (40+36+32+39+40)/195= 0.958 %.

4. Naïve Bayes Classifier with Normalized TF-IDF and statistical stemming N-Gram (N=2)

After using the test set in the program by technique two, the confusion matrix that is produced is shown in Table 4.7.

The overall classification accuracy achieved over all categories is (40+35+39+37+40)/195= 0.979 %.

	Agricul_ture	Art	Economics	Health_medicine	Sport	Total
Agriculture	**40**	0	1	0	0	40
Art	0	**35**	1	0	0	36
Economics	0	0	**39**	1	0	40
Health_medicine	2	0	0	**37**	0	39
Sport	0	0	0	0	**40**	40
Total	40	35	41	40	40	**195**

Fig 3 shows the summary of accuracy over the four techniques, so that the Naïve Bayes with N-Gram technique had the most accuracy while the Naïve Bayes with TF had the least accuracy. Fig 4 shows the average running time over the four techniques, Which shows that are the Naïve Bayes with TF technique had the shortest running time while the Naïve Bayes with N-Gram had the longest running time.

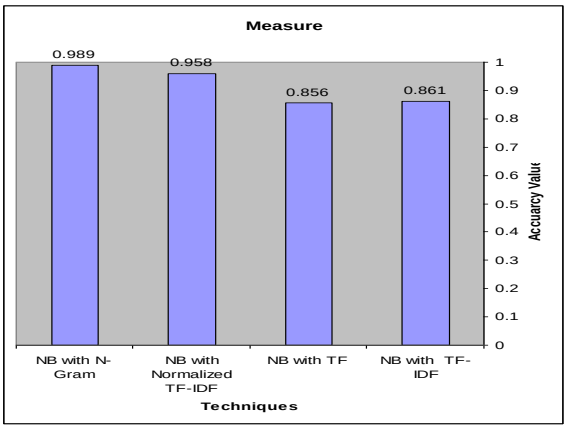

Figure 3 the accuracy for the four techniques

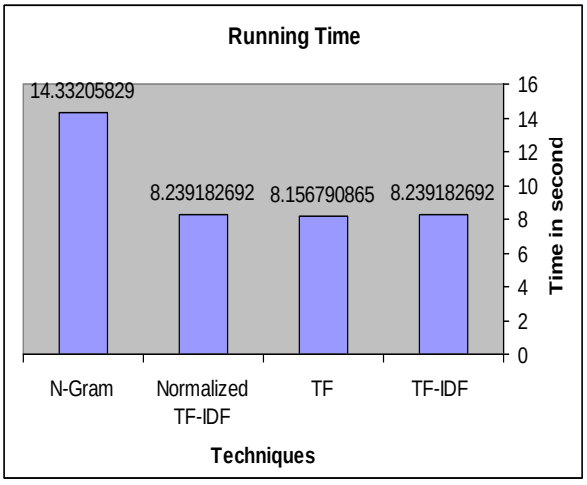

Figure 4 Running time for the four techniques

- Evaluation Result for the second test set

This section also will be divided into four subsections; each one will present the result of one of the four techniques

Table 7 : Confusion matrix for Hybrid approach of statisical N-Gram and F-IDF technique

1. Naïve Bayes with TF-IDF

After using the test set in the program by technique one, theoverall classification accuracy achieved over all categories can be calculated as (8+2+10+7+8)/49= 0.714 %.

2. Naïve Bayes Only

After using the test set in the program by technique two, the overall classification accuracy achieved over all categories is (0+1+10+8+8)/49= 0.551 %.

3. Naïve Bayes with Normalized TF-IDF

After using the test set in the program by technique two ,the overall classification accuracy achieved over all categories can be calculated as (8+8+8+6+9)/49= 0.795 %.

4. Naïve Bayes Classifier with Normalized TF-IDF and statistical stemming N-Gram (N=2)

After using the test set in the program by technique two , the overall classification accuracy achieved over all categories can be calculated as (8+9+8+8+10)/49= 0.877 %.

Table 8 show the summary of the accuracy result of the two test set depending on the four techniques, Fig 5 show the summary of the accuracy of the four techniques over the two test set.

Table 8: Summary of accuracy over the two test set

	NB only	TF-IDF	Normalized TF-IDF	Normalized Tf-IDF with N-Gram
Accuracy of first test set	0.856	0.861	0.958	0.979
Accuracy of second test set	0.551	0.714	0.795	0.877

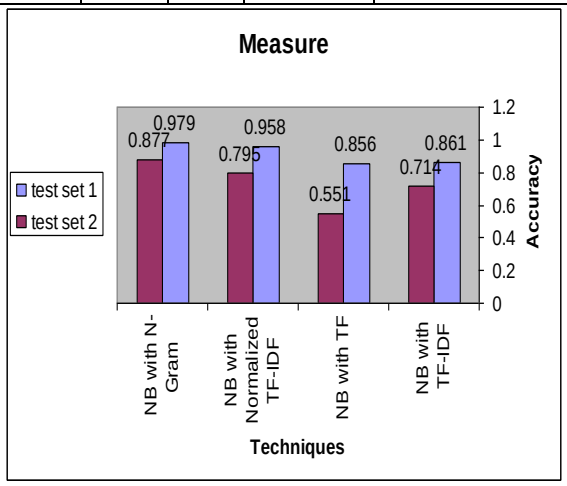

Figure 5 summary accuracy of the two tests set

V. CONCLUSION

This paper applied Naïve Bayes classifier algorithm for Arabic text categorization using four techniques.(TF,TF-IDF, Normalized TF-IDF and N-Gram with N=2 statistical stemmer with threshold similarity 0.8).The results showed that the four techniques differ in the accuracy and the running time. The four techniques are evaluated by two test set. The result showed that the Normalized TF-IDF and N-Gram with N=2 statistical stemmer with threshold similarity 0.8 technique is had the best accuracy.

The number of words in the category affects of the accuracy, and may lead to unfair categorization. Therefore, it is not suitable to use different length document without using normalized weight.

In addition the accuracy in [6] is closed to this accuracy result. Therefore, the hybrid approach of light and statistical N-Gram is the most suitable for Arabic text categorization when use Naïve Bayes classifier algorithm.

In the time criteria, the results are shown that the Naïve Bayes with N-Gram technique has the longest time, Vise versa the Naïve Bayes with only TF technique the shortest.

REFERENCES

[1] Robb,Drew, "Text Mining Tools Take on Unstructured data", Computerworld,2004.

[2] Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques",second edition, the morgan Kaufmann series in data management system, ISBN 1-55860-901-6, 2006

[3] Mark Dixon, "An Overview of Document Mining Technology", Citeseer,1997.

[4] Susan Dumais, John Platt and David Heckerman and Mehran Sahami, "Inductive Learning Algorithms and Representations for Text Categorization",CIKM98 : proceeding of the seventh international conference and information and knowledge management ,ACM.1998.

[5] Alaa M. El-Halees, "Arabic Text Classification Using Maximum Entropy", Citeseer, 2007.

[6] M. M. Syiam, Z. T. Fayed and M. B. Habib, "An intelligent system for Arabic text categorization", IJICIS, Vol.6, No. 1, JANUARY 2006

[7] http://en.wikipedia.org/wiki/Thomas_Bayes

[8] Loan Pop, "An approach of the Naïve Bayes classifier for the document classification" , General Mathematics Vol. 14, No. 4, pages 135-138, 2006.

[9] Ashraf Odeh, Aymen Abu-Errub, Qusai Shambour and Nidal Turab, "Arabic Text Categorization Algorithm Using Vector Evaluation Method", International Journal of Computer Science & Information Technology (IJCSIT) Vol 6, No 6, December 2014, pp 83-92.

[10] MajedIsmail Hussien, Fekry Olayah, Minwer AL-dwan & Ahlam Shamsan, "Arabic Text Classification Using Smo, Naïve Bayesian, J48 Algorithms", IJRRAS9 (2),November 2011, pp 306-319 , www.arpapress.com/Volumes/Vol9Issue2/IJRRAS_9_2_15.pdf

[11] Maria-Luiza Antonie and Osmar R. Zaïane, "Text Document Categorization by Term Association"'university of Alberta Canada,2002.

[12] http://www.scribd.com/doc/10552567/TFIDF

[13] http://www.computer.org/portal/web/csdl/doi/10.1109/CSSE.2008.829