Different Classification Algorithms Based on Arabic Text Classification: Feature Selection Comparative Study

Ghazi Raho
MIS Dept./ Amman Arab University
Amman-Jordan

Riyad Al-Shalabi
MIS Dept./ Amman Arab University
Amman-Jordan

Ghassan Kanaan
CS Dept./ Amman Arab University
Amman-Jordan

Asma'aNassar
CS Dept./ JUST University
Irbid-Jordan

***Abstract*—Feature selection is necessary for effective text classification. Dataset preprocessing is essential to make upright result and effective performance. This paper investigates the effectiveness of using feature selection. In this paper we have been compared the performance between different classifiers in different situations using feature selection with stemming, and without stemming.Evaluation used a BBC Arabic dataset, different classification algorithms such as decision tree (D.T), *K*-nearest neighbors (*K*NN), Naïve Bayesian (NB) method and Naïve Bayes Multinomial(NBM) classifier were used. The experimental results are presented in term of precision, recall, F-Measures, accuracy and time to build model.**

Keywords—Text Classification; Feature Selection; Arabic Text; Recall; F-Measure

I. INTRODUCTION

We know that the amount of Arabic information that founded on the internet is very large and increasing rapidly. This growth directs researchers to find some of the effectiveness mechanism and good tools that may help the researchers to better managing, filtering, processing and classification a large Arabic information resource. Text classification (TC) is the task using to classify a specific dataset into different classes; it also called document classification, text categorization or document categorization.

TC also used to solve some research problems such as information retrieval (IR), data mining, and natural language processing. There are many applications on TC like document indexing, document organization, text filtering, word sense disambiguation, speech recognition and web text hierarchical categorization.

TC can use as a binary classification like -nearest neighbors (KNN), Naïve Bayesian method and SVM and as a multi classification like boosting and multi-class SVM.

TC task can divides the dataset into two part: training set and testing set, the classifier algorithm learn on training to build a TC model, then TC system to classify the testing set into different classes, To achieve effective performance we used feature selection methods.

To get a better performance wedid some preprocessing steps on the dataset which we will talk about later in this paper. Section two will talk about the related work, section three will talk about our objectives, section four talk about experimental results, and then conclusion and future work, and finally the references.

II. RELATED WORK

In [1] the authors presented the performance of using a Support Vector Machines (SVMs) based text classification system on Arabic text. The authors using one of the feature selection methods which is CHI square method, theyuseda preprocessing stepsin their work to give a better evaluation. The proposed system gives good results. To classify any text we must determine a set of features to achieve best classification. This paper presents the effectiveness of six features selection method to extract and choose a good features from Arabic document. The authors used SVM classifier algorithm to compare the performance between these six methods (CHI, NGL, GSS, IG, OR and MI).

The authors in [2] used an in-house collected corpus from online Arabic newspaper archives, including Al-Jazeera, Al-Nahar, Al-Hayat, Al-Ahram, and Al-Dostor. The collected corpus consists of 1445 documents. These documents consist of nine categories, the authors did some Pre-processing for the dataset such as remove digits and punctuation marks, all the non-Arabic texts were filtered, remove the Arabic function words (stop words) and other. In [2] the result showing that CHI, NGL and GSS performed most effective with SVMs for Arabic TC tasks, but OR and MI performed terribly. In [3] the authors talked about three contributions: (i) showing successful classification of Arabic documents, (ii) make their database available to other researchers, (iii) find a better performancebetween Binary PSO and K-nearest neighbor using feature selection methods. In [3] the authors presented BPSO - KNN as a feature selection method and applied this method on three Arabic text dataset. The authors used three classification algorithms which are SVM, Naïve Bayes andC4.5 decision tree learning.

In [4] the authors used Chi-Squaremethod as a pre-processing step which applied on dataset before doing the classification. In [4] the authors compared between the proposed method and other feature selection methods the result shows that the proposed methodperformed better performance than other features selection methods.

III. Our Objectives

To compare the performance between different classification algorithm (decision tree, K-nearest neighbors(KNN), Naïve Bayesian method and Naïve Bayes multinomial classifier) in different situations: using feature selection methods with light stemmer, (khoja stemmer) and using feature selection with full word.

A. TC Process

Text classification system usually separated into three main phases which are : *Data preprocessing and feature selection phase that* makes the dataset more compatible and applicable to train the text classifier, *text classifier phase* that use to classify dataset into different classes, *and evaluation phase*to show the performance of the used classification algorithm.

B. Arabic Dataset Preprocessing

There are a lot of Arabic dataset available on the internet that can be used, we used BBC Arabic dataset that contains 4763 documents belongs to seven categories (News Middle East in 2356, News of the world in 1489, the economy and business 296, Sport 219, the press world 49, Science and Technology 232 Arts & Culture, 122). The dataset contains 1,860,786 words and 106,733key word. These dataset are processed according to the following steps:

1) Remove digits, dash, punctuation marks and any other mark.

2) Filtered all non-Arabic text.

3) Remove stop words from the text document (such as "بدا/", "أحد", "أخر" and other stop words).

4) Use feature selection methods with stemmer and with full word.

C. Feature Selection Methods

Feature selection (FS) is a task to choose a subset feature from the original feature set, FS is widely used in TC task. FS consist of following steps:

1) Feature generation: in this step we generate a subset of feature by using some search process.

2) Feature evaluation: in this step we used some evaluation matrices to measure the goodness of selected features.

3) Feature validation: in this step we used a validation procedure to measure if the selected features are valid or not.

In this paper we used two feature selection methods the Information Gain (IG), and theX^2statistics (CHI) as shown in table 1.

TABLE I. FS Methods

CHI	$\frac{N.[P(t_k,c_i).P(\bar{t}_k,\bar{c}_i)-P(t_k,\bar{c}_i)P(\bar{t}_k,c_i)]^2}{P(t_k)P(\bar{t}_k)P(c_i)P(\bar{c}_i)}$
IG	$\sum_{c\in\{c_i,\bar{c}_i\}}\sum_{t\in\{t_i,\bar{t}_i\}} P(t/c).log\frac{P(t/c)}{P(t).P(c)}$

D. Text Classifier

In this paper we used different classifiersthese classifiers are: decision tree, K-nearest neighbors (KNN), Naïve Bayesian method and Naïve Bayes multinomial,we have compared between the performance of these classifier in different terms of categorization effectiveness. wedivided the dataset into two parts, one for the training, and the other for testing.

E. TC Evaluation Measure

We have evaluated the performance for the classifiers (decision tree, K-nearest neighbors(KNN), Naïve Bayesian method and Naïve Bayes multinomial) in terms of precision, recall, accuracy, F-Measures and time to build model as shown in equations 1, 2, and 3.

$$Pi=TPi/(TPi+FPi) \quad (1)$$

$$Ri=TPi/(TPi+FNi) \quad (2)$$

$$Fi=2PiRi/(Ri+Pi) \quad (3)$$

IV. TC Experimental Results

We have used two feature selection methods (CHI and IG), four classifiers (decision tree, K-nearest neighbors(KNN), Naïve Bayesian method and Naïve Bayes multinomial classifier) were used, a Weka tools of version 3.7 were used, the results are shown in table II to table X.

TABLE II. Khoja Stemmer Experiments by Taking Chi-Square Feature Selection Type

Classifier type	Time to build model/ sec	Chi-Square feature selection results			
		accuracy	Average Precision	Average recall	F-Measures
D.T	33.67	99.6221 %	0.996	0.996	0.996
NB	4.01	90.9091 %	0.932	0.909	0.917
KNN	0.01	73.1262 %	0.807	0.731	0.716
NBM	0.16	92.7357 %	0.935	0.927	0.928

TABLE III. KHOJA STEMMER EXPERIMENTS BY TAKING IG RATIO SELECTION FEATURE

Classifier type	Time to build model	Info Gain			
		Accuracy	Precision	Recall	F-Measures
D.T.	36.27	99.6221	0.01	0.99	0.996
NB	4.61	90.9091	0.93	0.91	0.917
KNN	0.01	73.1262	0.81	0.73	0.716
NBM	0.06	92.7357	0.94	0.93	0.928

TABLE IV. KHOJA STEMMER EXPERIMENTS BY TAKING NO FEATURE SELECTION TYPE

Classifier type	Time to build model	Null Feature Selection Type			
		Accuracy	Precision	Recall	F-Measures
D.T.	31.5	99.4751 %	0.995	0.995	0.995
NB	4.26	90.9091 %	0.932	0.909	0.917
KNN	0.01	73.1262 %	0.807	0.731	0.716
NBM	0.17	92.7357 %	0.935	0.927	0.928

TABLE V. LIGHT STEMMER EXPERIMENTS BY TAKING CHI SQUARE FEATURE SELECTION TYPE

Classifier type	Time to build model	Chi-square Feature Selection Type			
		Accuracy	Precision	Recall	F-Measures
D.T.	49.4	99.4961 %	0.995	0.995	0.995
NB	5.43	91.9169 %	0.931	0.919	0.922
KNN	0.01	66.3657 %	0.891	0.664	0.675
NBM	0.17	92.0638 %	0.927	0.921	0.921

TABLE VI. LIGHT STEMMER EXPERIMENTS BY TAKING IG RATIO FEATURE SELECTION TYPE

Classifier type	Time to build model	Info-gain ratio Feature Selection Type			
		Accuracy	Precision	Recall	F-Measures
D.T.	54.1	99.5591 %	0.996	0.996	0.996
NB	7.07	91.9169 %	0.931	0.919	0.922
KNN	0.01	66.3657 %	0.891	0.664	0.675
NBM	0.06	92.0638 %	0.927	0.921	0.921

TABLE VII. LIGHT STEMMER EXPERIMENTS BY TAKING NO FEATURE SELECTION TYPE

Classifier type	Time to build model	Null Feature Selection Type			
		Accuracy	Precision	Recall	F-Measures
D.T.	44.05	99.5171 %	0.995	0.995	0.995
NB	6	91.9169 %	0.931	0.919	0.922
KNN	0	66.3657 %	0.891	0.664	0.675
NBM	0.07	92.0638 %	0.927	0.921	0.921

TABLE VIII. NULL STEMMER EXPERIMENTS BY TAKING CHI-SQUARE FEATURE SELECTION TYPE

Classifier type	Time to build model	Chi-square Feature Selection Type			
		Accuracy	Precision	Recall	F-Measures
D.T.	100.98	99.6221 %	0.996	0.996	0.996
NB	16.29	91.329 %	0.923	0.913	0.914
KNN	0.01	66.3867 %	0.781	0.664	0.63
NBM	0.05	92.0638 %	0.928	0.921	0.921

TABLE IX. NULL STEMMER EXPERIMENTS BY TAKING INFO GAIN RATIO FEATURE SELECTION TYPE

Classifier type	Time to build model	Info gain Feature Selection Type			
		Accuracy	Precision	Recall	F-Measures
D,T.	100.5	99.6221 %	0.996	0.996	0.996
NB	17.13	91.329 %	0.923	0.913	0.914
KNN	0	75.0577 %	0.802	0.751	0.734
NBM	0.13	92.0638 %	0.928	0.921	0.921

TABLE X. NULL STEMMER EXPERIMENTS BY TAKING NO FEATURE SELECTION TYPE

Classifier type	Time to build model	Null Feature Selection Type			
		Accuracy	Precision	Recall	F-Measures
D.T.	100.56	99.5801 %	0.996	0.996	0.996
NB	17.54	91.329 %	0.923	0.913	0.914
KNN	0	66.3867 %	0.781	0.664	0.63
NBM	0.2	92.0638 %	0.928	0.921	0.921

V. CONCLUSION

we have been investigated the performance of two FS methods with four classifiers(decision tree, K-nearest neighbors(KNN), Naïve Bayesian method and Naïve Bayes multinomial classifier) using Arabic dataset. The accuracy for decision tree, Naïve Bayesian method and Naïve Bayes multinomial better than K-nearest neighbors(KNN) in all cases. In Future work we will use more feature selection methods with different classifiers algorithms.

REFERENCES

[1] M. Abdelwad. "Chi Square Feature Extraction Based Svms Arabic Language Text Categorization System." *Journal of Computer Science*, 2007.

[2] M.Abdelwadood. "Support vector machines based Arabic language text classification system: feature selection comparative study." Advances in Computer and Information Sciences and Engineering.Springer Netherlands, 2008.

[3] C. Hamoudaand W. David. "Feature subset selection for Arabic document categorization using BPSO-KNN." Nature and Biologically Inspired Computing (NaBIC), Third World Congress on. IEEE, 2011.

[4] H. Bilal, A. Mansour, and Sh. Aljawarneh. "An Efficient Feature Selection Method for Arabic Text Classification." International Journal of Computer Applications ,2013.

[5] M.Abdulrahman, I. Hmeidi, and I. Alsmadi. "Indexing of Arabic documents automatically based on lexical analysis". arXiv preprint arXiv:1205.1602 2012.

[6] A. Saleh. "Automated Arabic Text Categorization Using SVM and NB."Int. Arab J. e-Technology 124-128. 2011.

[7] G, Sami, and N. Ben Amara. "Neural Networks and Support Vector Machines Classifiers for Writer Identification Using Arabic Script." International Arab Journal of Information Technology (IAJIT) 2008.

[8] R. Saleh, et al. "Bilingual experiments with an Arabic-English corpus for opinion mining. 2011.

[9] B, AlSalemi, , and M. Ab Aziz. "Statistical Bayesian Learning for Automatic Arabic Text Categorization." Journal of Computer Science 2011.

[10] A. Ahmed, H. Chen, and A. Salem. "Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums."ACM Transactions on Information Systems (TOIS)2008.