Naïve Bayesian Based on Chi Square to Categorize Arabic Data

Fadi Thabtah, Philadelphia University, Jordan, ffayez@philadelphia.edu.jo
Mohammad Ali H. Eljinini, AL-Isra Private University, Jordan, eljinini@ipu.edu.jo
Mannam Zamzeer, University of Jordan, Jordan, mzamzeer@ju.edu.jo
Wa'el Musa Hadi, AL-Isra Private University, Jordan, whadi@ipu.edu.jo

Abstract

Text classification is a supervised technique that uses labelled training data to learn the classification system and then automatically classifies the remaining text using the learned system. This paper investigates Naïve Bayesian algorithm based on Chi Square features selection method. The base of our comparisons are macro F1, macro recall and macro precision evaluation measures. The experimental results compared against different Arabic text categorization data sets provided evidence that feature selection often increases classification accuracy by removing rare terms.

Key Words: Text Categorization, Naïve Bayesian, Arabic Text Data, Chi Square.

1. Introduction

With the rapid growth of online information, Text Categorization (TC) has become one of the key techniques for handling and organizing text data. TC techniques are used to classify news stories, to end interesting information on the World Wide Web (WWW), and to guide a user's search through hypertext. Since building text classifiers by hand is difficult and time-consuming, it is advantageous to learn classifiers from examples. The goal of TC is the classification of documents into a fixed number of preened categories. Each document can be in multiple, exactly one, or no category at all. In this paper we focus on single label data sets.

Many TC approaches from data mining and machine learning exist such as Decision Trees [11], Support Vector Machine (SVM) [6], Rule Induction [10], Associative Classification [19], and Neural Network [22]. The goal of this paper is to present and compare results obtained against Arabic text collections using Naïve Bayesian (NB) algorithm.

The bases of our comparison of the NB are the most popular text evaluation measures (F1, Recall, and Precision) [21]. To the best of the author's knowledge, there are no comparisons which have been conducted against Arabic language data collections using NB algorithm based on Chi Square features selection methods.

The organization of this paper is as follows, related works are discussed in Section 2. TC problem is described in Section 3. In Section 4, experiment results are explained, and finally conclusions and future works are given in Section 5.

2. Related Works

Since TC stands at the cross junction to modern information retrevial and machine learning, several research papers have focused on it but each of which has concentrated on one or more issues related to such task. There are few previous works on Arabic TC.

For instance, [8] 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.

The author's of [14] presented results using statistical methods such as maximum entropy to cluster Arabic news articles; the results derived by these methods were promising without morphological analysis.

In [7], NB was applied to classify Arabic web data, the results showed that the average accuracy was 68.78%.

[5] 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 preprocessing techniques (normalization, stop words removal, and stemming).

The algorithm developed by [5] has outperformed other presented text classification algorithms, i.e. [7], [4], [14], and [12] Categorizer with regards to F-measure results.

[9] Used three classification algorithms, namely SVM, KNN and NB, to classify 1445 texts taken from online Arabic newspaper archives. The compiled texts were classified into nine classes: Computer, Economics, Education, Engineering, Law, Medicine, Politics, Religion and Sports. Chi Square statistics was used for feature selection. [9] Discussed that "Compared to other classification methods, our system shows a high

classification effectiveness for Arabic data set in terms of F-measure (F=88.11)".

In [18], the authors investigated different variations of Vector Space Model using KNN algorithm, these variations are Cosine coefficient, Dice coefficient and Jacaard coefficient, using different term weighting approaches. The average F1 results obtained against six Arabic data sets indicated that Dice based TF.IDF and Jaccard based TF.IDF outperformed Cosine based TF.IDF, Cosine based WIDF, Cosine based ITF, Cosine based log(1+tf), Dice based WIDF, Dice based ITF, Dice based log(1+tf), Jaccard based WIDF, Jaccard based ITF, and Jaccard based log(1+tf).

Finally, in [3] NB, and KNN were applied to classify Arabic text. The results show that the NB classifier outperformed KNN base on Cosine coefficient with regards to macro F1, macro recall and macro precision measures.

3. Text Categorization Problem

TC is the task in which texts are classified into one of the predefined categories based on their contents. If the texts are newspaper articles, categories could be, for example, Economics, Politics, Sports, and so on. This task has various applications such as automatic email classification and web-page categorization. Those applications are becoming increasingly important in today's information-oriented society.

TC problem can be defined according to [16] as follows: The documents divided in two datasets; one for training and one for testing. Let training data set = $\{d_1,d_2,\ldots,d_g\}$, where g documents are used as examples for the classifier, and must contain sufficient number of positive examples for all the categories involved. The testing data set $\{d_{g+1},d_{g+2},\ldots,d_n\}$ used to test the classifier effectiveness. The following matrix represents data splitting into training and testing parts, A document d_k is considered a positive example to C_y if C_{ky} =1 and a negative example if C_{ky} =0, See Table 1.

Generally, TC goes through three main steps: Data pre-processing, text classification and evaluation. Data pre-processing phase is to make the text documents suitable to train the classifier. Then, the text classifier is constructed and tuned using a text learning approach against from the training data set. Finally, the text classifier gets evaluated by some evaluation measures i.e. recall, precision, etc. The following sections are devoted to these three phases.

Table 1: Representation of text categorization problem

Category	Training data set			Testing data set		
	d_1	…	d_j	d_{j+1}	…	d_n
C_1	C_{11}	…	C_{1j}	$C_{1(j+1)}$		C_{1n}
…	…	…	…	…	…	…
C_m	C_{m1}	…	C_{mg}	$C_{m(j+1)}$	…	C_{mn}

3.1 Data Pre-processing

The data used in our experiments are The Sudia Press Agncey (SPA) data sets [20], SPA are collected from [15]. the data set consist of 1562 Arabic documents of different lengths that belongs to 6 categories, the categories are (Economic "اقتصادية" , Cultural "ثقافية" , Political " سياسية" , Social " إجتماعية" , Sports " رياضية " , General " عامة"), Table 2 represent the number of documents for each category.

Table 2: Number of Documents per Category

Category Name	Number of Documents
Cultural News	258
Sports News	255
Economic News	250
Social News	258
Political News	250
General News	255
Total	**1562**

Arabic text is different than English one since Arabic language is highly inflectional and derivational language which makes monophonical analysis a complex task. Also, in Arabic script, some of the vowels are represented by diacritics which usually left out in the text and it does use capitalization for proper nouns that creates ambiguity in the text. In this Arabic dataset, each document file was saved in a separate file within the corresponding category's directory, i.e. this dataset documents are single-labeled.

Representing Arabic dataset Documents: As mentioned before, this representation aims to transform the Arabic text documents to a form that is suitable for the classification algorithm. In this phase, we have followed [1],[2] [7] and processed the Arabic documents according to the following steps:

1. Each article in the Arabic data set is processed by removing the digits and punctuation marks.
2. We have followed [13] in the normalization of some of the Arabic letters such as the normalization of (hamza (ٕ) or (ٔ)) in all its forms to (alef (ا)).

3. All the non Arabic texts were filtered.
4. Arabic function words were removed.

The Arabic function words (stop words) are the words that are not useful in Information Retrieval systems e.g. The Arabic prefixes, pronouns, and prepositions. And to avoid high dimensionality we applied feature selection on Arabic data sets i.e. (Chi Square method). The following sub-section discusses Feature selection.

3.1.1 Feature Selection and Dimensionality Reduction
Feature selection is the process of selecting the best K terms as a subset of the terms occurring in the training set and using only this subset as features in TC.

Feature selection achieves two main goals. First, it makes the training applied to a classifier more efficient by decreasing the high dimensionality of effective vocabulary. Second, feature selection often increases classification accuracy by redaction rare term.

There are many feature selection methods such as Document Frequency (DF), Information Gain (IG), Chi-square Testing ($\chi 2$), and so on [23]. In this paper we focus on $\chi 2$ method. The following sub-section discuss $\chi 2$ method.

3.1.1.1 Chi-square Testing ($\chi 2$)
Chi-square testing ($\chi 2$) is a well-known discrete data hypothesis testing method from statistics, which evaluates the correlation between two variables and determines whether they are independent or correlated [17]. The test for independence, when applied to a population of subjects, determines whether they are positively correlated or not.

$\chi 2$ value for each term t in a category c can be defined by equation (1).

$$\chi 2\ (t,c) = \frac{N(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)} \qquad (1)$$

Where N: is the total number of training documents, A is the number of documents in c containing t, B is the number of documents not in c containing t, C is the number of documents in c not containing t, D is the number of documents not in c not containing t. $\chi 2$ was used in TC problem [23] and showed promising results.

3.2 Approaches to Text Categorization
There are many approaches to categorize text. This section covers NB approach. NB is a popular approach for TC. This section describes the general nature of the NB approach, its process for classifier training and document classification, and its advantages and disadvantages.

3.2.1 Naïve Bayesian
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 in TC problem we used equation 2.

$$p(class \mid document) = \frac{p(class).p(document \mid class)}{p(document)} \qquad (2)$$

Where:

P(class|document): It's 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 modelled as sets of words, thus the p(document|class) can be written like:

$$p(document \mid class) = \prod_i p(word_i \mid class) \qquad (3)$$

So:

$$p(class \mid document) = p(class) \prod_i p(word_i \mid class) \qquad (4)$$

Where:

P(wordi|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:

$$P(word_i \mid class) = (T_{ct} + \lambda)/(N_c + \lambda V) \qquad (5)$$

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.

4. Experiment Results
We used three evaluation measures (Recall, Precision, and F1) as the bases of our comparison, where F1 is computed based on the following equation:

$$F1 = \frac{2 * \Pr ecision * \mathrm{Re} call}{\mathrm{Re} call + \Pr ecision} \quad (6)$$

Precision and recall are widely used for evaluation measures in IR and ML, where according to Table 3,

$$\Pr ecision = \frac{a}{(a+b)} \quad (7)$$

$$\mathrm{Re} call = \frac{a}{(a+c)} \quad (8)$$

Table 3: Documents possible sets based on a query in IR

Iteration	Relevant	Irrelevant
Documents Retrieved	a	b
Documents not Retrieved	c	d

All of the NB experiments were implemented using VB.NET on 2.8 Pentium IV machine with 256 RAM.

In each experiment, a varying number of features were used, from 50 to 1000, depending on their $\chi 2$ values.

Figure 1 depicts the F1, Precision, and Recall results generated by the NB categorizer against six Arabic data sets; where in each data set we consider 70% of documents arbitrary for training, and 30% for testing.

After analyzing Figure 1, we found that the NB categorizer work well when the number of features decreased between 1000-800 features. NB reach a peak for F1 and precision when the number of features is 800 features, and reach a peak for recall measure when the number of features is 700 features.

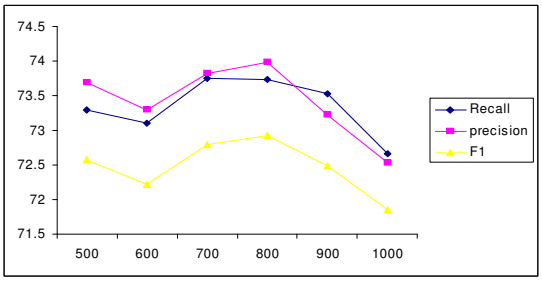

Figure 1:F1 measure, Precision and Recall results, Number of features varying from 500 to 1000

Figure 2 depicts the Macro F1, Precision, and Recall results generated by the NB categorizer.

After analyzing Figure 2, we found that the NB categorizer works well when the number of features grows. NB reaches a peak for Macro F1, Macro precision, and Macro Recall when the number of features is 50 features.

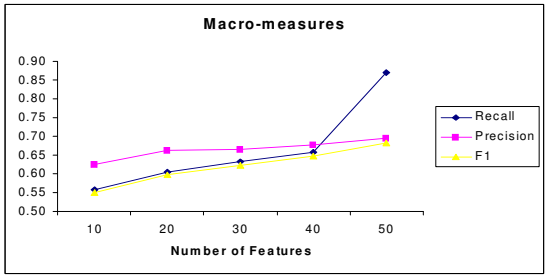

Figure 2: Macro F1 measure, Precision and Recall results, Number of features varying from 10 to 50

Figure 3 depicts the Macro F1, Precision, and Recall results generated by the NB categorizer, when the number of features varying from 100 to 500.

After analyzing Figure 3, we found that the NB categorizer worked well when the number of features increased, but dropped down when the number of features reached 500 features. NB reaches a peak for Macro F1, Macro precision, and Macro Recall when the number of features is 400 features. In general NB reaches a peak when the number of features is 800 features according macro F1 and macro precision, and reach a peck when the number of features is 50 features according macro recall measure.

Finally we conclude feature selection often increased classification accuracy by removing irrelevant terms.

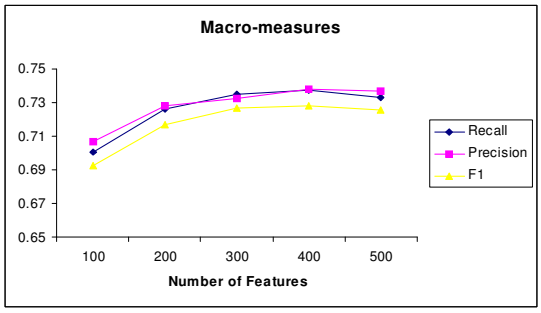

Figure 3: Macro F1 measure, Precision and Recall results, Number of features varying from 100 to 500

5. Conclusions and Future Works

In this paper we discussed the problem of automatically classifying Arabic text documents. We

used the NB algorithm which is based on probabilistic framework to handle our classification problem.
Feature selection often increases classification accuracy by redaction rare term. NB reaches a peak when the number of features is 800 features. In near future, we intend to survey more feature selection methods i.e.(DF, and IG) on [15]data sets.

References

[1] Benkhalifa, M., A. Mouradi, and H. Bouyakhf. "Integrating WordNet knowledge to supplement training data in semi-supervised agglomerative hierarchical clustering for text categorization," *Int. J. Intel Syst* (16:8), 2001, pp.929-947.

[2] Guo, G., H. Wang, D. Bell, Y. Bi, and K. Greer. "An kNN Model-based Approach and its Application in Text Categorization," In *proceedings of 5th International Conference on Intelligent Text Processing and Computational Linguistic, CICLing, LNCS 2945, Springer-Verlag,* 2004, pp.559-570.

[3] Hadi W., Thabtah F., ALHawari S., Ababneh J. "Naive Bayesian and K-Nearest Neighbour to Categorize Arabic Text Data ," *In proceedings of the European Simulation and Modeling Conference*, Le Havre, France, 2008.

[4] El-Halees A. "Mining Arabic Association Rules for Text Classification In the proceedings of the first international conference on Mathematical Sciences," *Al-Azhar University of Gaza, Palestine*, 15 -17 (2006).

[5] El-Halees A. "Arabic Text Classification Using Maximum Entropy The Islamic University," *Journal of Series of Natural Studies and Engineering* (15:1), 2007, pp.157-167.

[6] Joachims T. "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," *In Proceedings of the European Conference on Machine Learning (ECML),* 1998, pp.173-142, Berlin.

[7] El-Kourdi, M., Bensaid, A., and Rachidi, T. "Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm," *20th International Conference on Computational Linguistics*, 2004, Geneva.

[8] Laila K. "Arabic Text Classification Using N-Gram Frequency Statistics A Comparative Study,"*DMIN*, 2006, pp.78-82.

[9] Mesleh, A. A. "Chi Square Feature Extraction Based Svms Arabic Language Text Categorization System," *Journal of Computer Science* (3:6), 2007, pp. 430-435.

[10] Moulinier, I., Raskinis, G.,and Ganascia, J. "Text categorization: a symbolic approach," *In Proceedings of the Fifth Annual Symposium on Document Analysis and Information Retrieval*,1996.

[11] Quinlan, J. "C4.5: Programs for machine learning,". San Mateo, CA: Morgan Kaufmann, 1993.

[12] Sakhr software company's website: www.sakhrsoft.com, 2004.

[13] Samir, A., W. Ata, and N. Darwish. "A New Technique for Automatic Text Categorization for Arabic Documents," *5th IBIMA Conference (The internet & information technology in modern organizations),* 2005, Cairo, Egypt.

[14] Sawaf, H. Zaplo,J. and Ney. H. "Statistical Classification Methods for Arabic News Articles,". *Arabic Natural Language Processing,* Workshop on the ACL,2001. Toulouse, France.

[15] SPA website: www.spa.gov.sa

[16] F.Sebastiani, "A Tutorial on Automated Text Categorization*," In Proceedings of the ASAI-99, 1st Argentinian Symposium on Artificial Intelligence,* 1999. pp. 7-35.

[17] Snedecor, W., and Cochran, W. "Statistical Methods, Eighth Edition," Iowa State University Press.1989.

[18] Thabtah F., Hadi W., Al-shammare G., AlHawari S. "VSMs with K-Nearest Neighbour to Categorise Arabic Text Data," *To Appear in the Proceedings of the International Conference on Machine Learning and Data Analysis*, 2008, San Francisco, USA.

[19] Thabtah, F., Cowling, P., and Peng, Y. "MMAC: A new multi-class, multi-label associative classification approach," *In Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM)*, 2004, pp. 217-224, Brighton, UK

[20] Al-Thubaity,A. Almuhareb, A. Al-Harbi, S. Al-Rajeh, A. and Khorsheed, M." KACST Arabic Text Classification Project: Overview and Preliminary Results," *9th IBMIA Conference on Information Management in Modern Organizations,* 2008.

[21] Van Rijsbergan, C. "Information retrieval Buttersmiths," London, 2nd Edition, 1979.

[22] Wiener, E., Pedersen, J.O.,and Weigend, A.S.A neural network approach to topic spotting. Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval (SDAIR),1995, pp. 317-332,Las Vegas, Nevada.

[23] Yang, Y., and Pedersen, J.O. "A comparative study on feature selection in text categorization," *In Proc. of Int'l Conference on Machine Learning (ICML)*, 1997, pp. 412-420.