Automated Arabic Text Categorization Using SVM and NB

Saleh Alsaleem
Shaqra University, Saudi Arabia

Abstract: *Text classification is a supervised learning technique that uses labeled training data to derive a classification system (classifier) and then automatically classifies unlabelled text data using the derived classifier. In this paper, we investigate Naïve Bayesian method (NB) and Support Vector Machine algorithm (SVM) on different Arabic data sets. The bases of our comparison are the most popular text evaluation measures. The Experimental results against different Arabic text categorization data sets reveal that SVM algorithm outperforms the NB with regards to all measures*.

Keywords: *Text Categorization, Naïve Bayesian, Arabic Text Data, Support Vector Machine*

Received July 8, 2010; Accepted September 20, 2010

1. Introduction

Text categorization (TC) is one of the important tasks in information retrieval and data mining. The problem of TC has been active for four decades, and recently attracted many researchers due to the large amount of documents available on the World Wide Web, in emails and in digital libraries. Automated TC involves assigning text documents in a test data collection to one or more of the pre-defined classes/categories based on their content. Unlike manual classification, which consumes time and requires high accuracy, Automated TC makes the classification process fast and more efficient since it automatically categorizes documents.

The goal of TC task is to assign class labels to unlabelled text documents from a fixed number of known categories. Each document can be in multiple, exactly one, or no category at all. In this paper we focus on just a single label assignment.

Many TC approaches from data mining and machine learning exist such as: decision trees [14], Support Vector Machine (SVM) [9], rule induction [13], and Neural Network [24]. The goal of this paper is to present and compare results obtained against Saudi Newspapers (SNP) Arabic text collections [1] using Support Vector Machine (SVM) algorithm and Naïve Bayesian (NB) algorithm. The bases of our comparison of the SVM and NB are the most popular text evaluation measures (F1, Recall, and Precision) [21]. In other words, we want to determine the categorizer that produces the best results. To the best of the author's knowledge, there are no comparisons which have been conducted against SNP data collections using NB and SVM and evaluated using Recall, F1 and Precision measures.

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 retrieval 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. In last decade, there are many previous works conducted on Arabic TC.

For instance, [11] 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 [17] 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 [5], NB was applied to classify Arabic web data; the results showed that the average accuracy was 68.78%.

[3] 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).

The algorithm developed by [4] has outperformed other presented text classification algorithms, i.e. [5], [3, 17, 5] Categorizer with regards to F-measure results.

[12] 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. [12] 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 [20], 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).

In [7], 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 show that the NB classifier outperformed KNN base on Cosine coefficient with regards to macro F1, macro recall and macro precision measures.

Finally, in [19] investigate NB algorithm based on Chi Square features selection method. The experimental results compared against different Arabic text categorization data sets provided evidence that feature selection often increases classification accuracy by removing rare terms.

3. Text Categorization Problem

TC, also known as text classification, is the task of automatically sorting a set of documents into categories (or classes, or topics) from a predefined set. Such task is related to IR and ML communities. Automated text classification tools are attractive since they free organizations from the need of manual categorization of document, which can be too expensive, or simply not feasible given the constraints of the application or the number of documents involved [18].

TC involves many applications such as automated indexing of scientific articles according to predefined thesauri of technical terms, filing patents into patent directories, selective dissemination of information to information consumers, automated population of hierarchical catalogues of web resources, spam filtering, identification of document genre, authorship attribution, survey coding and even automated essay grading.

TC problem can be defined according to [18] as follows: let G denote the collection of categories which contain {g_1,g_2,…..,g_n}, let D denote the collection of documents and Q is an incoming text. Also, let R denote the set of classifiers for $D \times Q \rightarrow G$, each document d ε D is assigned a single class g that belongs to G. The goal is to find a classifier h ε H that maximizes the probability that r(d) = G for each test case (d, g).

Generally, TC task goes through three mainly steps: Data pre-processing, text classification and evaluation. Data preprocessing 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, precisinon, etc . The next two sub-sections are devoted to discuss the main phases of the TC problem related to the data we utilised in this paper.

3.1. Data Pre-Processing on Arabic Data

The data used in our experiments are The Saudi Newspapers (SNP) [1], the data set consist of 5121 Arabic documents of different lengths that belongs to 7 categories, the categories are (Culture "الثقافية " , Economics "الإقتصادية" , General "العامة" , Information Technology " تكنولوجيا المعلومات " , Politics " السياسية", Social " الأجتماعية ", Sport " الرياضة"), Table 1 represent the number of documents for each category.

Arabic text is different from 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 capitalisation for proper nouns that creates ambiguity in the text [8]. In the Arabic data set we use, each document file was saved in a separate file within the corresponding category's directory.

Moreover, we represented the Arabic data set to a form that is suitable for the classification algorithm. In this phase, we have followed [2, 6, 5] data format and processed the Arabic documents according to the following steps:

- Each article in the Arabic data set is processed to remove the digits and punctuation marks.
- We have followed [16] in the normalization of some Arabic letters such as the normalization of (hamza (إ) or (أ)) in all its forms to (alef (ا)).
- All the non Arabic texts were filtered.
- Arabic function words were removed. The Arabic function words (stop words) are the words that are not useful in IR systems e.g. The Arabic prefixes, pronouns, and prepositions.

3.2. Approaches to Text Categorization

This section covers two existing approaches to text categorization: SVM, and NB. SVM is one of the effective algorithms that perform classification by constructing an N-dimensional hyperplane that optimally separates the data into two categories. NB which is a simple probabilistic classifier based on Baye's theorem. The next two subsections describe the general nature, process for classifier training and document classification, advantages and disadvantages, of both learning methods that we consider.

Table 1. Number of documents per category.

Category Name	Number of Documents
Culture	738
Economics	739
General	728
Information Technology	728
Politics	726
Social	731
Sport	731
Total	**5121**

3.2.1. Support Vector Machine

SVM was introduced by [22] as a class of supervised machine learning techniques. It is based on the principle of structural risk minimisation. In linear classification, SVM creates a hyper plane that separates the data into two sets with the maximum-margin. A hyper plane with the maximum-margin has the distances from the hyper plane to points when the two sides are equal. Mathematically, SVMs learn the sign function $f(x) = \text{sign}(wx + b)$, where w is a weighted vector in R^n. SVMs find the hyper plane $y = wx + b$ by separating the space R^n into two half-spaces with the maximum-margin. Linear SVMs can be generalised for non-linear problems. To do so, the data is mapped into another space H and we perform the linear SVM algorithm over this new space. SVM has been successfully used on TC [9,10] and they derived better results than other machine learning techniques such as NB, decision trees, and KNN with reference to accuracy.

3.2.2. 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 on the TC problem we use equation 1.

$$\text{p(class} | \text{ document)} = \frac{\text{p(class).p(document} | \text{class)}}{\text{p(document)}} \quad (1)$$

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:

$$p(document | \ class) = \prod_i p(wordi | \ class) \quad (2)$$

So

$$p(class | \ document) = p(class) \prod_i p(wordi | \ class) \quad (3)$$

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:

$$\text{P(wordi|class)= (Tct +λ)/(Nc+ λV)} \quad (4)$$

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.

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:

$$F1 = \frac{2 * \Pr ecision * \text{Re}\,call}{\text{Re}\,call + \Pr ecision} \quad (5)$$

Precision and recall are widely used evaluation measures in IR and ML, where according to Table 2,

$$\Pr ecision = \frac{a}{(a+b)} \quad (6)$$

$$\text{Re}\,call = \frac{a}{(a+c)} \quad (7)$$

To explain precision and recall, let's say someone has 5 blue and 7 red tickets in a set and he submitted a query to retrieve the blue ones. If he retrieves 6 tickets where 4 of them are blue and 2 that are red, it means that he got 4 out of 5 blue (1 false negative) and 2 red (2 false positives). Based on these results, precision=4/6 (4 blue out of 6 retrieved tickets), and recall= 4/5 (4 blue out of 5 in the initial set).

Table 2. Documents possible sets based on a query in IR.

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

Table 3 gives the F1, Recall, and Precision results generated by the two categorizers (NB, and) against SNP data sets where in each data set using ten-fold cross-validation.

Table 3. Results F1, Recall, and Precision of Arabic text categorization

Category Name	SVM			NB		
Evaluation measures	Precision	Recall	F1	Precision	Recall	F1
Culture	0.706	0.747	0.726	0.723	0.747	0.735
Economics	0.86	0.907	0.883	0.879	0.834	0.856
General	0.517	0.512	0.514	0.439	0.397	0.417
Information Technology	0.907	0.86	0.883	0.86	0.838	0.849
Politics	0.865	0.853	0.859	0.81	0.901	0.853
Social	0.633	0.617	0.625	0.505	0.547	0.525
Sport	0.961	0.949	0.955	0.968	0.917	0.942
Average	0.779	0.778	0.778	0.741	0.74	0.74

Cross validation is a known evaluation method in data mining, where the training data is divided randomly into n blocks, each block is held out once, and the classifier is trained on the remaining n-1 blocks; then its error rate is evaluated on the holdout block. Therefore, the learning procedure is executed n times on slightly different training data sets.

All the experiments were conducted using The Weka open source software [23].

After analysing Table 3, we found that the SVM categorizer outperformed NB on six data sets with regards to F1 results. Precision results obtain that the SVM outperformed NB on four data sets and NB outperformed SVM on three data sets.

Also Recall results obtain that the SVM outperformed NB on five data sets, NB won on a single data set and tied on one data set. The average of three measures obtained against seven Arabic data sets indicated that the SVM algorithm dominant NB algorithm.

Moreover, and for the Islamic data sets, the SVM classifier have 3.8%, 3.8% and 3.8% higher Recall, Precision and F1 figures than NB respectively.

Another notable result that was also reported is that all measures vary among categories. For example, the "Sport" category has a neat classification F1 of 94.2%, while the "General" category has a noticeably poor Recall of 41.7% using NB. These poor results indicate that the "General" category is highly overlapped with other categories.

Finally, SVM and NB classifiers perform excellent in the SNP data sets. This excellent result may indicate that the classes in the Islamic data sets are well defined and that there are a number of distinctive terms associated with each class in the data sets.

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 and SVM algorithm to handle our classification problem.

The average of three measures obtained against SNP Arabic data sets indicated that the SVM algorithm outperformed NB algorithm regards to F1, Recall and Precision measures. In near future, we intend to propose a new multi-label classification approach based on association rule for the text categorization problem.

References

[1] Al-Harbi S., Almuhareb A., A. Al-Thubaity, M. S. Khorsheed, and A. Al-Rajeh. Automatic Arabic Text Classification. JADT 2008: 9es Journées Internationales d'Analyse Statistique des Données Textuelles. (pp. 77-83).

[2] 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.

[3] 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).

[4] 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.

[5] 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.

[6] 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.

[7] Hadi W., Thabtah F., ALHawari S., Ababneh J. (2008b) 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), 2008.

[8] Hammo, B., Abu-Salem, H., Lytinen, S., and Evens, M. 2002. "QARAB: A Question Answering System to Support the Arabic Language". Workshop on Computational

Approaches to Semitic Languages. ACL 2002, Philadelphia, PA, July. pp. 55-65.

[9] Joachims T. (1999). Transductive Inference for Text Classification using Support Vector Machines. Proceedings of the International Conference on Machine Learning (ICML), (pp. 200-209). 1999.

[10] 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.

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

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

[13] 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.

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

[15] Sakhr Software Company’s website: www.sakhrsoft.com, 2004.

[16] 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.

[17] 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.

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

[19] Thabtah F., Eljinini M., Zamzeer M., Hadi W. (2009) Naïve Bayesian based on Chi Square to Categorize Arabic Data. In proceedings of The 11th International Business Information Management Association Conference (IBIMA) Conference on Innovation and Knowledge Management in Twin Track Economies, Cairo, Egypt 4 - 6 January. (pp. 930-935).

[20] Thabtah F., Hadi W., Al-shammare G. (2008) VSMs with K-Nearest Neighbour to Categorise Arabic Text Data. In The World Congress on Engineering and Computer Science 2008. (pp.778-781), 22-44 October 2008. San Francisco, USA.

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

[22] Vapnik V. (1995). The Nature of Statistical Learning Theory, chapter 5. Springer-Verlag, New York.

[23] WEKA. Data Mining Software in Java: http://www.cs.waikato.ac.nz/ml/weka, 2001.

[24] 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.

Saleh Alsaleem Received his Bachelor of Education in science in field of Computer Science from King Saud University, Riyadh, Saudi Arabia 1991, Master of Science, MS, in computer science from Ball State University, Muncie, Indiana, USA, in July 1996, and Doctor of Philosophy (Ph.D.) in Computer Science from Wayne State University, Michigan, USA, in May 2001. He worked as faculty member in department of computer technology at Riyadh College of Technology and then head of the department in the same college. After that he worked as program coordinator in Arab Open University. Currently he is working as the dean of Admission and Registration and acting as general supervisor for IT services in Shaqra University.