Arabic Text Mining Using Maximum Entropy Model

[1]Al-Gaphari.G., [2]Al-Nuzaili A
[1]Department of Computer Science
Faculty of Computers & Technology
The University of Sana'a. Yeman
drghalebh@yahoo.com

[2]Department of Mathematics
Faculty of Science
The University of Sana'a. Yeman

ABSTRACT: *Building a high performance classifier requires an efficient training algorithm as well as a high performance testing algorithm .With the present effort, we propose to focus on the development of an automated Maintainable information classifying system as a main goal. The system has two phases: learning phase and testing phase. On the one hand, the system accepts a set of Arabic classified documents as a real training data set, during its learning phase. The system learning technique based on the so- called Maximum Entropy Model. The model enables the system to learn the parameters weights. On the other hand, the system accepts, during its testing phase, any randomly selected Arabic unclassified document or documents. Then it uses the estimated weights, learned so far, to decide whether that document or documents belongs or belong to one or more predefined categories, depend on their context. The maximum entropy model was implemented. Hence, making a conclusion that emphasizes the model efficiency for Arabic text categorization in terms of learning speed, real time classification speed, and classification accuracy.*

Keywords: Text mining, Text classification, Arabic text processing, Maximum entropy model

Received: 11 January 2010, Revised 1 March 2010, Accepted 5 March 2010

1. Introduction

An automatic text mining is to concern with the task of automatic extracting relevant information, from natural language text, and to search for interesting relationships between the extracted entities [1]. Automatic text classification is one of the basic techniques in the area of automatic text mining. It is one of the more difficult data-mining problems, since it deals with very high-dimensional data sets with arbitrary patterns of missing data [2]. An automatic text categorization is to automatically assign a natural language text to one or more predefined categories based on their content. Automatic text categorization plays an important role in many applications: authorship attribution, information organization, word sense disambiguation, Hierarchical Categorization of web pages and management tasks [25].

In fact, the more the information volume available on the Internet and corporate Intranets lasts to increase, the more the growing interest in helping people better find, filter, and manage that resources. The most widespread automatic text categorization application is used for assigning subject categories to documents to support text retrieval and filtering [3].

Automatic categorization technologies must be able to help classification structures. Such structures are very general, consistent through individuals, and relatively static. For example Google's topic hierarchy, as well as those which are more dynamic and customized to individual interest, such as specific conferences emails [4].

Some web search engines companies employee trained professionals to classify new elements [15].This process is very time

consuming and costly, also, it has a limitation in its capability. Therefore, there are increased demands in developing technologies for automatic Arabic text classification [14, 15]. In fact, there are several research projects to investigate and find out the techniques in automatically classifying English documents as well as other languages. Also, there are some software have been developed for English text classification such as Construe system which used for classifying Reuters news stories in nineteen's [14]. Some other software developed for the same purpose relay on inductive learning mechanism which uses labeled data set for training. It is a matter of fact; text classification poses many challenges for inductive learning techniques because there might be a huge number of term features. Fortunately, the resulting classifiers have many advantages where they allow users to tradeoff precision and recall based on their tasks. There is an increasing number of machine learning techniques have been implemented to text classification [8, 22].

Unfortunately, there is a limitation in both research and software development in terms of automatic Arabic text classification.The main objective of this paper is to describe results of software classifier implementation ,where The maximum entropy model is used as a supervised learning technique in build a such classifier. The classifier is trained and evaluated on The AFP Arabic newswire corpus. The corpora contain newswire stories and provided by the Lingustic Data Consertium (LDC) .

2. Literature Review

There is a limited number of research papers for Arabic text categorization ,the most closely related work to ours are survayed and reported :

S. Al-Harbi et al . 2008 [1] attemped to obtain a better understanding and elaboration of Arabic text classification technique using SVM and C5.0 classification algorithms. The C5.0 classifier outperformed the SVM classifier by about 10%, the SVM average accuracy reported was 68.65% while the C5.0 average accuracy reported was 78.42%

Jihong Cai and Fei Song 2008 [2] stated that they explored the use of different feature selection methods for text categorization using maximum entropy modeling. They also proposed a new feature selection method based on the difference between the relative document frequencies of a feature for both relevant and irrelevant classes. They reported that their experiments on the Reuters RCV1 data set show that their own feature selection performs better than the other feature selection methods. And maximum entropy modeling is a competitive method for text categorization.

Fadi Thabtah et al. 2008[3] invisitigated different variations of VSM using KNN algorithm,they mentioned in their paper the variations that are implemented in the experiment. Such variations are cosine coeffecient,Diece coeffecient and Jacaard coeffecient . They concluded that the Dice based TF-IDF and Jaccard based TF-IDF outperformed Cosine coeffecient approach with regards to F1 results.

Mohammed Naji Al-Kabi and Saja I. Al- Sinjilawi 2007 [4] presented a paper that described the design and the implementation of a new suitable method for classifying Prophet Mohammed's (PBUH) .They compared six classifying techniques in their experiment. They reported that Naïve Bayesian classifier is the first best among other classifiers with 85% of accuracy.

Rehab Duwairi 2007[5] presented a paper that compares the performance of three classifiers for Arabic text categorization, such classifiers are Naïve Bayes, K-nearest-neighbors and distance based classifier. The researcher stated that she represented the documents as a vector of words. Then the stemming was performed. Finally, the researcher concluded that Naïve Bayes classifier outperforms the other tow.

Alaa Al-Halees 2007[6] proposed a structure for a classifier system which includes five parts. The researcher stated that he implemented a part of the system called ArabCat. ,it increases the f-measure from 68.13% to 80.41% .

Reda A.El-Khboribi and Mohammed Ismael 2006[7] proposed a paper that described Arabic text categorization system development. They stated that system was based on statistical learning. They concluded that the system was powerful in terms of grasping the semantic of documents so that it has promising results.

Victor Chan et al. 2006 [8] reported that they created a computer system that is able to predict earthquakes. They stated that they applied the concept of data mining to gather data on earthquakes and examine them. Because there is so much data on earthquakes, they had decided to focus on the Bay Area. Thus, they had used data located in this region for input. The results they found would be used as a guide to determine future earthquakes [21].

Address Hotho et al 2005[9] gave a brief introduction to the field of text mining; also they gave a brief overview of current available text mining techniques, their features and their application to specific problems. They concluded that their studies shown a rough overview of the text mining field and several starting points for further studies.

Un Yong Nahm and Raymond J. Mooney2002 [10] presented a framework for text mining, called Disco TEX (Discovery from Text Extraction); they used a learned information extraction system to transform text into more structured data that can be mined for interesting relationships. They concluded that text-mining systems can be developed relatively rapidly and evaluated easily, on existing IE corpora, by utilizing existing IE and data mining technology.

Joshua Goodman, 2002 [11] described a speedup for training conditional maximum entropy models, they stated that the algorithm they used was a simple version on Generalized Iterative Scaling ,but converges roughly an order of magnitude faster, depending on the number of constraints and the way speed is measured.

Kamal Nigam et al. 1999[12] proposed a paper that described the use of maximum entropy techniques for text classification. They stated that Maximum entropy is a probability distribution estimation technique widely used for a variety of natural language tasks. They concluded that maximum entropy is used for text classification by estimating the conditional distribution of the class variable given the document.

Kostas Fargoes et al. [16] proposed a weighted maximum entropy model (WMEM) for text classification. Their work used a feature selection strategy and assigning weights to the features with the χ^2-test. They evaluated WMEM performance over 10 categories of the Reuters 21578 dataset. Where the average accuracy obtained was 70.64%. Finally, they concluded that WMEM performs better than the other classifiers in the 'money-fx'.

Maximum entropy probability models introduce a clear mean to combine diverse pieces of contextual evidence in order to estimate the probability of a certain linguistic class occurring with a certain linguistic context [16]. In this experiment, we first demonstrate how to represent evidence, then how to implement maximum entropy as an inductive learning algorithm, for Arabic text classification.

3. Facts Representations

The facts are represented with functions called contextual predicates and features. If C = $\{c_1, c_2, \ldots, c_n\}$ represents the set of possible categories that will be predicted [8]. D Represents the set of possible contexts that can be observed, and then a contextual predicate is a function: $\lambda : D$ {true, false} which returns a value true or false based on the presence or absence of useful information in some context $d \in D$. The set of contextual predicates of each problem is provided by experimenter. They usually are used in features as in the following figure:

$$\text{f: C X D} \longrightarrow \{0,1\}$$

$$f_{\lambda,\alpha}(c,d) = \begin{cases} 1 & \text{if c = } \alpha \text{ and d = true} \\ 0 & otherwise \end{cases} \quad (1)$$

$$f_{(s,b)}(d,c) = \begin{cases} 1, & if\ c = s, b\ occurs\ in\ d \\ 0, & otherwise \end{cases} \quad (2)$$

Figure 1. Facts representations

And they make sure for the co-existence of some predictions á with some contextual predicates λ

4. Feature Selection using χ^2 Test

When implementing machine learning techniques for text categorization, it is very important to start up with feature selection process. Unfortunately, among the most challenging jobs in the categorization process is to make a right selection of suitable features to represent a particular class instance [14]. In fact, selection of the best candidate features may be a real disadvantage for the selection algorithm, in terms of effort and time consuming. In this experiment the binary representation is implemented, which means a term either appears or does not appear in the text of interest.The χ^2 test is curried out to decrease the high dimensionality of the text and for the weighting purposes of maximum entropy framework.

The χ^2 static can be given as $\chi^2 = \sum \frac{(observed - expected)^2}{expected}$ (3)

Where $\sum$ stands for summation and computed over the classes of possible outcomes.

Let's select an example of our experiment implementation data set, which is Linguistic Data Consortium (LDC), data set.Assume that there are two different classes c_1="Economy "and $c_2 \neq$" Economy "and we are interested in estimating the Independence of the term "direction" with classes c_1 and c_2 by computing the term frequencies of the term "direction" in the training data set, we find that the term "direction"occurs with class c_1="Economy"69 times, with other classes that is not the $c_2 \neq$" class Economy "35 times.

The total term frequencies in class c_1=" Economy "is 415119 terms, while in the other classes a total of 584881, which is equivalent to the total of N=1000000 terms overall in the data set. Hence the probabilities of class c_1=" Economy "and the term "direction" can be calculated as follows:

P (c_1="Economy") =415119/1000000 and P (term="direction") =69/1000000.

The hypothesis about the independence (null hypothesis) is that existence of the term=" direction "and the class" Economy "are independent, therefore, the null hypothesis can be computed as follows:

H_0: P ("direction","Economy ") =p (c_1="Economy")*p (term="direction").

Then the χ^2 is computed by (3), the critical value for a significance level á=0.05 is found for one degree of freedom, if the computed value is greater than the critical value then the null hypothesis that the term "direction"and the class "Economy" occur independently. Therefore, for each large computed χ^2 value there exist a strong evidence for the pair ("direction "," Economy ").The term "direction" is a good feature for the categorization in the class" Economy" .The main objective of this work is to build a classifier based on maximum entropy model. In this experiment, we start selecting the most representative features among the large numbers of nominees and perform categorization in a lower dimensionality space.

5. Maximum Entropy Modeling

Maximum entropy model is a mechnism for learning probability distribution from data.It widely used for a variety of natural language tasks,such as language modeling and text classifications[17]. The basic principle of maximum entropy is to prefer uniform distributions in terms of there is no external knowledge. Constraints imposed on the distribution that derived from labeled training data inform the mechanism where to be minimally non-uniform [18, 23].

The data for a classification problem is described as a number of features [4]. These features can be quite complex and allow the experimenter to make use of prior knowledge about what types of information are expected to be important for classification [19]. Each feature corresponds to a constraint on the model [13]. The maximum entropy approach has a unique solution which can be obtained by the improved iterative scaling algorithm. In another word

$$H(x) = -\sum_{x \epsilon \omega} p(x) \log (p(x))$$

Where x = (a, b), a $\in$ C (a set of categories), b $\in$ D (a set of documents) ù= C x D .The correct distribution p(x) is that maximizes entropy, subject to the constraints setup by the experimenter [24]. The log linear model

$$p(\vec{x}, c) = \frac{1}{z} \prod_{i=1}^{k} \propto^{f_i(\vec{x}, c)} \quad (4)$$

$$\text{Where } z = \sum_{\vec{x}, c} \prod_{i=1}^{k+1} \propto_i^{f_i(\vec{x}, c)} \quad (5)$$

And k is the number of features and α is the weight for feature $\mathbf{f_i}$ and **c** ranges over all possible

Classes or categories [17]. The Generalized iterative scaling algorithm guarantees to find the maximum entropy distribution **p*** that satisfies the constraint $Ep^* f_i = Ep^{\sim} f_i$ by adding one additional feature (not binary) to meet the requirement of the algorithm [21].

$$f_{k+1}(x) = C - \sum_{i=1}^{k} f_i(x) \quad (6)$$

$$\text{Where } C = \text{constant (max } (\sum_{i=1}^{k} f_i(x) \quad) \quad (7)$$

The iterative scaling algorithm is shown in Figure 2.

6. Maximum Entropy Classifier Software

As was stated earlier, the objective of this paper is to build maximum entropy classifier software (MECS) based on maximum

entropy model and to evaluate the performance and the efficiency of that classifier in terms of Arabic text categorization. Therefore, a computer program was constructed which accepts a set of Arabic classified documents as a training data set that can be loaded to the program during its learning phase, and then the program accepts one or more than one Arabic unclassified

1. Initialize $\{\propto_i^{(1)}\}$ =1 while i=1, 2, k+1, compute

$Ep\tilde{}f_i=\frac{1}{N}\sum_{j=1}^{N} f_i(\vec{x},c)$ Set n=1.

2. Compute $p^{(n)}(\vec{x}, c)$ for the distribution $p^{(n)}$ given by $\{\propto_i^{(n)}\}$ for each element ($\vec{x}$, c) in the training set:

$$p^{(n)}(\vec{x}, c) = \frac{1}{z}\prod_{i=1}^{k+1}(\propto_i^{(n)})^{f_i(\vec{x},c)}$$

where $z = \sum_{\vec{x},c}\prod_{i=1}^{k+1}(\propto_i^{(n)})^{f_i(\vec{x},c)}$

3. Compute $Ep^{(n)}f_i$ for all i=1, 2… k+1 by

$$Ep^{(n)}f_i=\frac{1}{N}\sum_{j=1}^{N}\sum_{c} p(c|\vec{x}_j) f_i(\vec{x}_j,c) \quad (12)$$

4. Update the parameters $\propto_i$ by

$$\propto_i^{(n+1)}=\propto_i^{n}\left(\frac{E_p - f_i}{E_{p^{(n)}} f_i}\right)^{\frac{1}{C}}$$

If the parameters of the procedure have converged, stop else increment n and go to 2.

Figure 2. The algorithm of maximum entropy model.

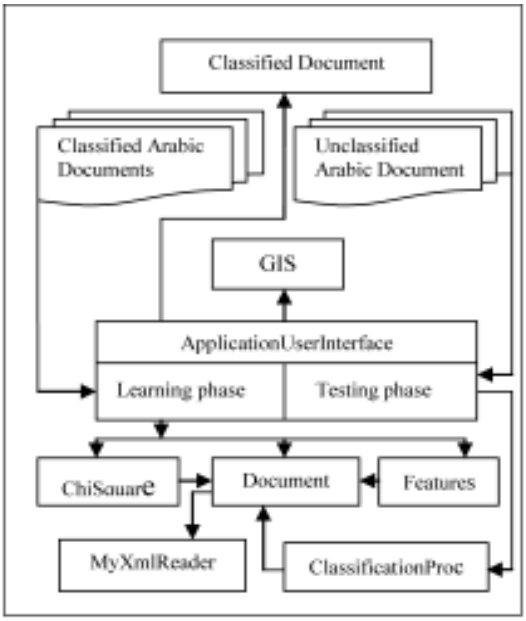

Figure 3. Program hierarchy for maximum entropy model.

documents as testing data set which can be loaded to the program during its testing phase .Hence, the program returns the category/ categories contain that document / documents.

The program hierarchy is shown in Figure 3.

The software system comprises of seven software modules, each one in turn consists of some subcomponents. Each subcomponent performs some individual tasks. Then such tasks are integrated with the other tasks created by other subcomponent within the first component. The same process is repeated within the remaining components; hence, the ultimate task integration takes place within the whole software system.

6.1. Corpus

Each document in the training data set is assigned to one of the predefined categories. There are different resources for research in machine learning and text categorization. Reuters-21578, Reuters-21450 and Reuters-810000 text categorization test collection are very popular and typical examples for English text classification [12]. The Linguistic Data Consortium (LDC) provides two Arabic collections, the Arabic GIGAWORD and the Arabic NEWSWIRE-a corpus. The program read 10,535 training documents and 2,250 testing documents of the AFR newswire-a, all together made more than 12,000,000 terms. Such documents classified into seven categories, those categories are economy, politics, sport, technology, arts, health and culture.

Arabic stop words	Meanings
فى	How
بين	Between, within
اين	Where
من	From
مع	With

Table 1. Arabic stop words

6.2. Experiment Results

The software starts up with documents preprocessing .It starts filtering the content of the documents by removing any numeric digit, non-word character, white space and any stop words .Table 1 shows some Arabic stop words with their English translation.

Then the software computes the χ^2 static based on the following equation .

$$x^2 = \frac{(a11*a22-a21*a12)^2}{(a11+a12)*(a11+a21)*(a12+a22)*(a21+a22)} \quad (8)$$

Some output results of this software can be shown in Table 2.

Arabic terms	English translation	χ^2 values
الشركة	The Company	36.0447
الفائدة	The interest	29.6307
صناعة	An Industry	25.3610
البنتخب	League	27.4830
نقاط	Points/scores	25.1440
الأمة	The Nation	86.4080

Table 2: Arabic terms with thei χ^2 values .

The next step of the software is to select, for the maximum entropy model, the most 1500 higher ranked terms for each class. Table 3 shows some classes with the 5 top ranked terms based on their χ^2 test.

6.3. Performance Evaluation and Validation

The total number of documents used in this experiment is 12785. They are divided into two sets: training set and validation set. The training set includes 10535 documents while the validation set includes 2250 documents. The performance of the classifier is measured over the validation set. For each tested set of documents, the classification accuracy is computed by the following measure $(a+d)/(a+b+c+d)$, also, some other measures are used, [21, 23], such as

Precision $=a/(a+b)$, Recall= $a/(a+c)$ and Fall Out= $b/(b+d)$

where a is the number of documents in the category of interest that are correctly assigned to the category, b is the number of documents in the category of interest that are incorrectly assigned to the category, while c is the number of documents are not in the category of interest that are incorrectly assigned to the category and d is the number of documents are not in the category of interest that are correctly assigned to the category [20,22]. In addition to above measures, ***F***-measure is used

and given by:

$$F = (2 * p * r)/(p + r) \quad (9)$$

Category	Recall	Precision	Breakeven Point%
اقتصاد Economy	95%	95.00%	100 %
سياسة Politics	100%	72.22%	83.87%
رياضة Sport	92.85%	100.00 %	96.29%
تكنولوجيا Technology	90.28%	89.95%	90.11%
فنون Arts	88.44%	78.98%	83.44%
صحة Health	91.89%	92.97%	92.43%
ثقافة Culture	87.76%	85.55%	86.64%

Table 3. 5 Top ranked terms by the χ^2 test

Where p is the classifier precision and r is the classifier recall. As stated above, the maximum entropy classifier performance is calculated by the above measures and shown over seven categories Table 4. In fact, the obtained recall of the classifier is 0.9223 while the obtained precision is 0.8781; consequently, the classifier breakeven is 0.90397. Thus, the maximum entropy classifier outperforms all those classifiers reported in previous studies. In fact, it gives promising result.

Category	Recall	Precision	Breakeven Point%
اقتصاد Economy	95%	95.00%	100 %
سياسة Politics	100%	72.22%	83.87%
رياضة Sport	92.85%	100.00 %	96.29%
تكنولوجيا Technology	90.28%	89.95%	90.11%
فنون Arts	88.44%	78.98%	83.44%
صحة Health	91.89%	92.97%	92.43%
ثقافة Culture	87.76%	85.55%	86.64%

Table 4. Maximum Entropy Classifier Performance

7. Summary and Conclusion

In this paper the maximum entropy model was used for Arabic text categorization. A software classifier was designed and constructed for implementing that model. The dataset was represented in binary representation. The right selection of suitable features among candidate features was made by chi-square technique for the binary representation. A real-world dataset was used for testing and validating the software classifier performance, the result of the experiment was very promising .It shown that the maximum entropy classifier (MECS)outperforms other classifiers used for text categorization, so

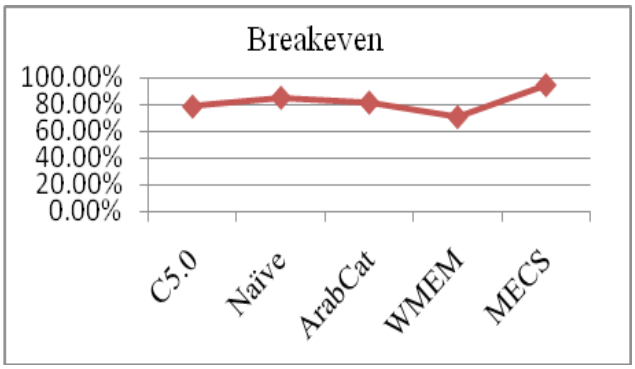

Figure 4. Classifiers with their performance

far, such as Naïve Bayesian, Cosine ,C5.0 and WMEM. In fact, the maximum entropy classifier in this experiment raised the performance average, from 85%, as reported by some researchers in [1], [4], [6], [16] up to 90.4 % as obtained in this experiment. Also, the comparison result is shown in Table 4 and Figure 4.

The future work is going to deal with Arabic text classification using different representations and different features selections techniques, to reduce the dimensionality, such as Sebastiani. Also, advanced methods of machine learning will be selected as well as other optimization techniques, such as genetic algorithms, association rule, and knapsack. The comparative study among selected algorithms will take place. The LDC dataset will be used as training and testing datasets the performance will be measured over the validation dataset.

8. Acknowledgement

The author would like to thank anonymous reviewers for their useful comments.

References

[1] Al-Harbi, S., Al-muhareb, A., Al-Thubaity, A., Khorsheed, M. S., Al-Rajeh, A (2008). Automatic Arabic Text Classification, *In*: JADT 9th Juornees Internationales d'Analyse statistique des Donnees Textuelles, p.77-83.

[2] Cai, J., Song, F (2008). Maximum Entropy Modeling with Feature Selection for Text Categorization. Springer-Verlag Berlin Heidelberg. p.1-10

[3] Thabtah, F., Hadi, W. M., Al-shammare, G.(2008). VSMs with K-Nearest Neighbour to Categoriese Arabic Text Data, *In*: Proceedings of the World Congress on Engineering and Computer Science 2008,WCECS,San Francisco, USA. p. 1-4.

[4] Al-Kabi, M., N., Al- Sinjilawi, S.I. (2007). A comparative Study of the Efficiency of Different Measures to Classify Arabic Text, *University of Sharjah Journal of Pure & Applied Sciences*, 4 (2) 13-26.

[5] Duwairi, R. (2007). Arabic Text Categorization, *The International Arab Journal of Information Technology.* 4 (2) 125-131

[6] El-Hlees, A. (2007). Arabic Text Classification Using Maximum Entropy, Elsevier, p.158-167.

[7] El-Khboribi, R., A., Ismail, M., A (2006). An Intelligent System Based on Statistical Learning for Searching in Arabic Text, *AIML Journal*, 3 .41-47.

[8] Chan, V., Cheung, K., Lee, B. McVay, K Truong, T. Vu, L., (2006). TT Interactive Data Mining Division Earth Quake Analysis and Predictions, San Francisco. p.1-10.

[9] Hotho, A. Nurnberger , A., Paab, G (2005). A Brief Survey of Text Mining, University of Kassel.

[10] Nahm, U. Y., Mooney, R. J. (2002). Text Mining with Information Extraction, AAAI Technical Report SS. USA. p.60-67.

[11] Goodman, J.(2002). Sequential Conditional Generalized Iterative Scaling, *In*: Proceedings of the 40th Annual Meeting of

the Association for Computational Linguistics Philadelphia . p. 9-16.

[12] Nigam, K., Kachites, A. Thruny, S., Mitchelly T. (2000). Text Classifications from Labeled and Unlabeled Documents using EM. Kluwer Academic Publishers, Boston. Manufactured in the Netherlands. p.1-34.

[13] Grossman, D.,Frieder, O. (2004). Information Retrieval Algorithms and Heuristics, Springer, USA.

[14]Manning, C., D., Schutze, H. (2000). Foundations of Statistical Natural Language Processing" The MIT Press Cambridge, Massachusetts London, England Second printing, Massachusetts Institute of Technology.

[15] Dumais, S., Platt, J., Heckerman, D., Stanford , M (2000). Inductive Learning Algorithms and Representations for Text Categorization "Microsoft Research and Stanford University, USA. p.1-8

[16] Fragos , K. Maistors, Y., Skourlas , C (2001). A weighted Maximum Entropy Language Model for Text Classification, University of Athens Press, p.1-9. http://nts.ece.ntua.gr/nlp_lab.

[17] Lu, Q., Getoor, L. (2003). Link-based Text Classification, University of Maryland, p.1-7.

[18] Malouf , R. (2002). A comparison of algorithms for maximum entropy parameter estimation, The Netherlands. p.1-7.

[19] Ratnaparkhi , A. (1998). Maximum Entropy Models for Natural Language Ambiguity Resolution, The University of Pennsylvania press, p.1-147.

[20] Beeferman, D., Berger A., Lafferty, J. (1997). Text Segmentation Using Exponential Models, http://arxiv.org/abs/cmp-lg

[21] Berger, A. L., Pietra S., Pietra, V (1996). A maximum Entropy Approach to Natural Language Processing, Association for Computational Linguistics. p.1-71.

[22] Smirnov, E. (2005). Personal communication. June, 15, 2005.

[23] Han. J., Kamber, M (2006). Data Mining Concepts and Techniques, Elsevier, London.

[24] Witten, I., H., Frank, E. (2006). Data Mining Practical Machine Learning Tools and Techniques, Elsevier, London.

[25] Kantardzic, M. (2003). Data Mining: Concepts, Models, Methods, and Algorithms, John Wiley & Sons.