ISSN 2277-3061

# Effects of Classification Techniques on Medical Reports Classification

Elfadil A. Mohamed[1], Fathi H. Saad[2], Omer I. E. Mohamed[3]
[1]College of Business Administration, Al Ain University of Science and Technology, UAE
elfadil.mohammed@aau.ac.ae
[2]NHS Oxfordshire, Oxford, UK
fathi.saad@nhs.net
[2]Department of Computer Science, Alkhawarizmi College, UAE
omareldai@gmail.com

## ABSTRACT

Text classification is the process of assigning pre-defined category labels to documents based on what a classifications has learned from training examples. This paper investigates the partially supervised classification approach in the medical field. The approaches that have been evaluated include Rocchio, Naïve Bayesian (NB), Spy, Support vector machine (SVM), and Expectation Maximization (EM). A combination of these methods has been conducted. The experimental result showed that the combination which uses EM in step 2 is always produces better results than those uses SVM using small set of training samples. We also found that reducing the features based on tf-tdf values is decreasing the classification performance dramatically. Moreover, reducing the features based on their frequencies improve the classification performance significantly while also increasing efficiency, but it may require some experimentation

### INDEXING TERMS/KEYWORDS

Document classification, positive-class based learning, partially supervised classification, labelled and unlabeled data, medical text mining, and features reduction

ISSN 2277-3061

## 1. INTRODUCTION

Classification is a form of data analysis that extracts models describing important data classes [10]. The extracted models are called classifiers which are used to predict categorical class labels. The medical field has recently received great attention regarding the analysis of medical data which is available in an electronic form. The nature of the medical data is either unstructured or semi-structured which make it difficult to be analyzed using traditional data mining techniques. The medical staffs need automatic classification methods to analyze and categorize this huge amount of data. The Gastroenterology unit of a local hospital in UK had just such a problem as they collected electronic reports on thousands of colonoscopy procedures, but could not give answer to simple questions, such as the percentage of successful colonoscopies undertaken [34]. The aim of colonoscopy is to check for medical problems such as bleeding, colon cancer, polyps, colitis, etc. [6].

Text classification is a two-step process, consisting of a learning step and classification step. When the class label of each training data is provided, this is called supervised learning. The supervised classification has recently being used in the medical domain for small number of training sample data; see for example [15]. However, supervised learning has the problem of the considerable effort required to manually label a large number of training examples for every class, particularly for multi-class problems. As supervised learning methods, most existing text classification algorithms require sufficient training examples so that the obtained classification model produces accurate results [22]. When the number of training examples in each class decreases, the classification accuracy of traditional text classification algorithms degrade dramatically. In medical domain, this is serious problem, because labelled documents are often very sparse because manually labelling data is boring, tiring, costly and continuing for a long time. To solve this problem, exploiting unlabeled documents in text classification has become an active research problem in text classification recently. This leaded to a new approach called partially-supervised classification. There have been a number of techniques reported in developing partially-supervised text classification recently. Most of these techniques require labelled training examples of all classes. It has been reported that those techniques obtain considerable improvement over traditional supervised techniques when the size of training examples is small [22]. So partially supervised approach is very attractive for the medical domain since medical practitioners are often very busy dealing with patients and cannot be expected to spend large amounts of time labelling data.

The approach we used in this paper has recently been introduced [1, 2] for binary text classification problems. Further evaluation of the approach can be found recently in [34]. This study is based on the use of a large set of unlabeled documents and a small set of labelled documents for every class so as to reduce the labelling efforts. The approach we used took this idea further and uses only positive and unlabeled documents to learn the classifier based on theoretical study reported in [35], cutting down more on the labelling effort. So this technique is different from the other partially-supervised classification techniques as it doesn't require labelling of negative training examples. This approach is two-step strategy. Step 1 identifies the positive documents from the unlabeled documents, and step 2 builds the final classifier. There are a number of algorithms that are applicable in step 1 and step 2. Deciding on what algorithms should be applied is not a trivial task, but is required for the effective application of the technique to real-world data.

The main purpose of this paper is to perform a comprehensive practical evaluation of partially supervised classification technique approaches to classify real-world medical reports. The approaches that will be investigated are Rocchio (Roc), NB and Spy for step 1, and SVM and EM for stem 2. The methods available in each step of the process will be tested in combination. The combination that produces the best performance according to some evaluation measures will be recommended. The evaluation will be performed through a real-world medical problem: the classification of a set of colonoscopy reports. For further efficiency, we will also experiment on reducing the set of features used to represent a document.

The rest of the paper is organized as follows: reviewing related works in section 2. The methods and algorithms used by partially supervised classification approach explained in section 3. In section 4 we described the data set used. Document representation is explained in section 5. The performance measures used are presented in section 6. The document pre-processing is outlined in section 7. Our methodology is described in section 8. We presented and analyzed the results in section 9. Finally, section 10 concludes the paper.

## 2. RELATED WORK

The labelled documents availability problem resulted in new research direction focuses on exploiting the unlabelled documents in text classification. Here is a brief description of some related methods that use unlabelled examples to improve text classification. Co-Training approach [22], splits the feature set by $\mathbf{x}=(\mathbf{x}^1, \mathbf{x}^2)$ and trains two classifiers $\theta_1$ and $\theta_2$ each of which is sufficient for classification. The algorithm initially constructs two classifiers based on labelled data, and mutually selects several confident examples to expand the training set. The approach assumes that the two feature sets are conditional independent. Similar approach reported in [24], this approach instead of using two conditional independent features it co-trained two SVM classifiers using two feature spaces. One is the original feature space and the other is derived from clustering the labelled and unlabeled data. Another co-training based approach [25] used two hybrid algorithms, co-EM and self-training, using two randomly split features in co-training setting. More reports use co-training approach could be found in [26, 27].

Transductive Support Vector Machine (TSVM) approach [23], it maximizes margin over both the labelled data and the unlabelled data. It works by finding a labelling of the unlabeled data $D_u$ and a hyperplane which separates both $D_l$ and $D_u$ with maximum margin. Another reported works improving classification using unlabelled examples [28, 29, 30]. All above methods use small labelled examples for every class and large unlabelled examples for learning improve the classification. In the case of binary classification small labelled sets for both positive and negative classes are required. More recent approach focuses on the binary text classification problem. This approach combined the advantages of partially supervised classification approach which uses small set of labelled examples for every class and large unlabelled examples, and the advantages of theoretical study [10] which requires labelling small set of positive class only. Some studies called this approach positive class based learning [1, 13], but still in many reports known as partially supervised classification. This is two-step approach; first step is identifying a set of reliable negative documents, the second step is build the classifier using the labelled positive and identified negative documents. There are many methods applied different approaches for both steps. One technique called PEBL [31] identifying strong reliable negative documents using method says "strong negative documents are those documents that do not contain any features of the positive data". After a set of strong negative documents is identified, SVM is applied iteratively to build a classifier. Another techniques called S-EM [2] uses new method called 'spy' for the first step to identify reliable negative document, and uses EM for the second step. Reference [13] reports another technique called Roc-SVM which uses Rocchio algorithm for step 1 and SVM for step 2. In [32] one-class SVM is proposed; this technique uses only positive examples to build a SVM classifier. More details about methods used for step 1 and step 2 will be discussed in partially supervised classification section.

## 3. PARTIALLY SUPERVISED CLASSIFICATION

As we mentioned earlier, the approach we used in this paper is partially supervised classification. This approach uses a reduced set of positive documents, P, and a large set of unlabeled documents, U. There is initially no labelling of negative documents. The first step of the text classification is therefore to identify a reliable set of negative documents, RN, from the unlabeled documents. In this section we are describing the techniques that we are going to evaluate for both steps 1 and 2. The algorithms that we are going to use for step one are:

Naïve Bayesian classifier (NB) – NB is a popular classification technique and has been reported as performing extremely well in practice for text classification [12]. In NB, the document is considered an ordered list of words. The vocabulary is the set of all words considered for classification. To perform classification, NB compute the posterior probability $Pr(c_j|d_i)$, where $c_j$ is a class and $d_i$ is a document. In classifying a document $d_i$, the class with the highest $Pr(c_j|d_i)$ is assigned as the class of the document. Identifying a set RN of reliable negative documents from the unlabeled set U is done as follows in Figure 1.

Rocchio (ROC) – Rocchio classifiers are well explained in [11]. In this technique, each document d is represented as a vector, and each element in this vector represents a word. Each word value is calculated using tf-idf scheme [20]. The unlabelled set U is treated as negative set in this technique, then the positive set P and U used as the training data to build a Rocchio classifier which is used to classify U. the algorithm that use Rocchio to identify a set RN of reliable negative documents from U is the same as that in Figure 1 except that Rocchio classifier used instead of NB.

Spy technique – this technique is introduced in [2]. The name reflects the fact that some documents randomly selected from P are added to U and act as spies from the positive set, P, to the unlabeled set, U. They form the spy set, S, and behave like unknown positive documents in U. The spy set, S, allows the classifier algorithm to infer the behavior of the unknown positive documents in U. It then runs EM algorithm using the set $P-S$ as positive and the set $U \cup S$ as negative. After EM completes, a threshold t is employed to make the decision. Those documents in U with lower probabilities $Pr(c_j|d_i)$ than t are the most likely negative documents RN. Those documents in U (spies are not included) that have higher probabilities than t become unlabeled documents U. The reader is referred to [2] for details. Figure 2 bellow shows spy algorithm in S-EM. Step 2, iteratively applying a classification algorithm to the newly labelled data. Since some documents are still in the unlabeled set, U- RN, the chosen classifier is applied repeatedly to the data with the intention of extracting more possible negative data at each iteration and improving the overall performance of the classifier. The procedure will stop when no further negative documents are found in the unlabeled set, U-RN. There are two classifiers within this step that will be tested: Expectation-Maximization (EM) and Support Vector Machines (SVM).

Expectation-Maximization (EM) - This algorithm iterative algorithm for maximum likelihood estimation in problems with missing data [16, 19]. It consists of two steps, the Expectation step, and the Maximization step. The Expectation step basically fills in the missing data. In our case, it produces and revises the probabilistic labels of the documents in U-RN. The parameters are estimated in the Maximization step after the missing data are filled. EM converges when its parameters stabilize. Referring to EM algorithm shown in Figure 3 bellow, using NB in each iteration, EM employs the same equations as those used in building a NB classifier (line 3 for the Expectation step, and lines 1 and 2 for the Maximization step). The class probability given to each document takes the value between 0 and 1. Basically, EM iteratively runs NB to revise the probabilistic label of each document in set Q = U-RN. Since each iteration of EM produces a NB classifier, S-EM also has a mechanism to select a good classifier [2, 17].

1. Assign each document in *P* the class label 1;
2. Assign each document in *U* the class label -1;
3. Build a NB classifier using *P* and *U*;
4. Use the classifier to classify *U*. Those documents in *U* that are classified as negative form the reliable negative set *RN*.

Figure 1. The NB method for Step 1

1. *RN* = *NULL*;
2. *S* = *Sample*(*P*, *s*%);
3. *Us* = *U* ∪ S;
4. *Ps* = *P*-*S*;
5. Assign each document in *Ps* the class label 1;
6. Assign each document in *Us* the class label -1;
7. EM (*Us*, *Ps*); // This produces a NB classifier.
8. Classify each document in *Us* using the NB classifier;
9. Determine a probability threshold *t* using *S*;
10. **for** each document *d* ∪ *Us*
11. **if** its probability *Pr*(1|*d*) < *t* **then**
12. *RN* = *RN* ∪ {*d*};

Figure 2. The algorithm of Spy technique in S-EM.

1. Each document in P is assigned the class label 1;
2. Each document in RN is assigned the class label -1;
3. Each document d ∈ Q (= U-RN) is not assigned any label initially. At the end of the first iteration of EM, it will be assigned a probabilistic label, Pr(1|d). In subsequent iterations, the set Q will participate in EM with its newly assigned probabilistic classes.
4. Run the EM algorithm using the document sets, P, RN and Q until it converges.

Figure 3. The EM algorithm with the NB classifier.

Support Vector Machines (SVMs) – SVMs are linear functions of the form $f(x)=\mathbf{w}^T\mathbf{x} + b$, where $\mathbf{w}^T\mathbf{x}$ is the inner product between the weight vector **w** and the input vector **x**. SVM selects a hyperplane that separates the positive and negative examples while maximizing the smallest margin [1]. Since small set of RN is identified in step 1, SVM uses P, RN and U-RN and runs iteratively to build the classifiers. In each iteration new RN will be identified and added to RN set to build new classifier. The iteration converges when no document in U-RN classified as negative. The final classifier will be selected from the set of classifiers. The reader is referred to [13] for a detailed description of the SVM algorithm.

ISSN 2277-3061

## 4. DATA SET

IN THIS PAPER, WE FOCUSED ON THE ANALYSIS OF COLONOSCOPY PROCEDURES, SINCE THIS IS THE MOST PRESSING TASK FROM A MEDICAL POINT OF VIEW, GIVEN THE INTRODUCTION OF COLORECTAL SCREENING IN THE UK IN 2006. COLONOSCOPY REFERS TO THE PASSAGE OF THE COLONOSCOPE FROM THE LOWEST PART (ANUS AND RECTUM) RIGHT AROUND THE COLON TO THE CAECUM AND IN SOME CASES INTO THE TERMINAL ILEUM VIA THE ILEO-CAECAL VALVE. THE AIM OF COLONOSCOPY IS TO CHECK FOR MEDICAL PROBLEMS SUCH AS BLEEDING, COLON CANCER, POLYPS, COLITIS, ETC. AFTER EACH COLONOSCOPY PROCEDURE, THE ENDOSCOPIST AT THE NNUH WOULD GENERATE A DETAILED REPORT ABOUT THE CURRENT STATUS OF THE EXAMINED PART OF THE BODY AND THE RESULT OF THE PROCEDURE ITSELF USING A SOFTWARE PACKAGE CALLED ENDOSCRIBE. CLASSIFYING THE PROCEDURE AS SUCCESSFUL OR FAILED DEPENDS ON WHAT IS WRITTEN IN THE REPORT.

The dataset used contained 4,876 documents for as many colonoscopy procedures. 25% of these documents were selected using 1-in-4 include sampling strategy to be used as test documents. The rest (75%) were used to create training sets. We used the dataset in two ways. The first way focuses on classifying the reports into successful or failed procedure. To achieve this, we combined extensive database querying, looking for regular expressions, with an expert doctor to create the "gold-standard" classification against which to measure our automatic classification. The "gold-standard" classification assigned 656 documents to the class of failed colonoscopies and 4,220 documents to the class of successful colonoscopies, giving an overall failure rate of 13.45%. This way will be used to evaluate all the approaches under investigation. Table 1 shows the class distribution for each set of documents, according to the "gold-standard" classification.

TABLE I

THE "GOLD-STANDARD" SUCCESSFUL/FAILED CLASS DISTRIBUTION

| Data set | Successful | Failed | Total |
|---|---|---|---|
| Train | 3,298 | 359 | 3657 (75%) |
| Test | 1,042 | 177 | 1219 (25%) |
| Total | 4,220 (86.5%) | 656 (13.5%) | 4876 (100%) |

The second way is looking at the data from different medical point of view which is diagnosis view. In this way we focused on certain diagnosis, namely "Diverticulosis", "Polyps", "Sessile Polyps" and "Ulcerative Colitis". We used this way for further investigation of the top three approaches that obtained good results using the first way. Using the same data in different ways as same as we use different datasets, because the class in each way is completely different. Creating the "gold-standard" classification is similar to the one we used in the first way. Tables 2, 3, 4 and 5 show the class distribution for Diverticulosis, Polyps, Sessile Polyps, and Ulcerative Colitis set of documents, according to the "gold-standard" classification. The "Yes" column in these tables refers to the positive class and "No" column refers to the negative class for that diagnosis. Figure 4 shows the numbers and percentages of the positive class for the 5 classification problems. The x axe shows the 5 classification problem, the left and right y axes show the number and percentage of the positive class respectively.

TABLE 2

THE "GOLD-STANDARD" POSITIVE/NEGATIVE DIVERTICULOSIS CLASS DISTRIBUTION

| Data set | Yes | No | Total |
|---|---|---|---|
| Train | 687 | 2970 | 3657 (75%) |
| Test | 223 | 996 | 1219 (25%) |
| Total | 910 (18.7%) | 3966 (81.3%) | 4876 (100%) |

TABLE 3

THE "GOLD-STANDARD" POSITIVE/NEGATIVE POLYPS CLASS DISTRIBUTION

| Data set | Yes | No | Total |
|---|---|---|---|
| Train | 1178 | 2479 | 3657 (75%) |
| Test | 375 | 844 | 1219 (25%) |
| Total | 1553 (31.8%) | 3323 (86.2%) | 4876 (100%) |

TABLE 4

THE "GOLD-STANDARD" POSITIVE/NEGATIVE SESSILE POLYPS CLASS DISTRIBUTION

| Data set | Yes | No | Total |
|---|---|---|---|
| Train | 908 | 2749 | 3657 (75%) |
| Test | 282 | 937 | 1219 (25%) |
| Total | 1190 (24.4%) | 3686 (75.6%) | 4876 (100%) |

TABLE 5

THE "GOLD-STANDARD" POSITIVE/NEGATIVE ULCERATIVE COLITIS CLASS DISTRIBUTION

| Data set | Yes | No | Total |
|---|---|---|---|
| Train | 235 | 3422 | 3657 (75%) |
| Test | 79 | 1140 | 1219 (25%) |
| Total | 314 (6.4%) | 4562 (93.6%) | 4876 (100%) |

## 5. TEXT REPRESENTATION

The experiments reported in this paper also address the problem of which features to be included in the classification process. The kind of linguistic features used in this paper to represent documents are single words. Single words are the structural units of language made up of one individual term [5]. The most frequently used method to represent text is bag-of-words representation where all words from the set of documents are taken and no ordering of words or any structure of text is used [4]. Each distinct word corresponds to a feature of the set of documents. Each feature can either have a Boolean value to indicate the presence or absence of the term in the document, or the term frequency (TF) can be used as the feature value instead. Alternatively, the term frequency-inverse document frequency (tf-idf) [20] which is refined model of (TF). In our experiments we used tf-idf.

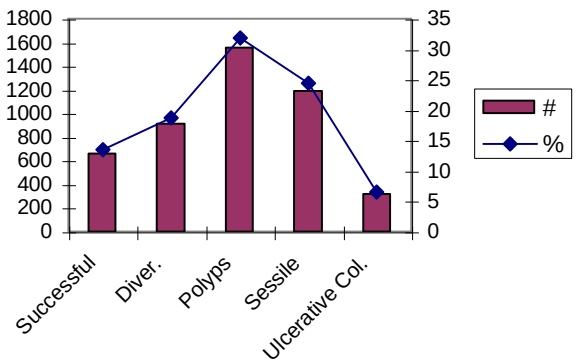


Figure 4. The Numbers and Percentages of the Positive Class for the 5 Classification Problems

ISSN 2277-3061

## 6. PERFORMANCE MEASURES

Four different measures were used to evaluate the performance of different classifiers: precision, recall, F-measure and accuracy [14]. Consider the following confusion matrix shown in Table 6 to help define these measures.

Precision is the percentage of correctly identified positive documents over those classified as positive (Equation 1). Recall is the percentage of correctly identified positive documents over all positive documents (Equation 2). Accuracy is the ratio of correct classification for the overall document set (Equation 3). The classifier that assigns class $C_+$ to all documents may have 100% recall but unacceptably low precision. Conversely, if the classifier did not assign any document to class $C_+$ it could have a perfect precision but low recall. The F-measure has been proposed to balance recall and precision by giving them equal weights (Equation 4). Therefore, for the evaluation of text classifiers, precision and recall need to be used in conjunction with the F-measure and/or accuracy.

Precision = a / (a+b), (1)

if (a+b) > 0 otherwise Precision = 1

Recall = a / ( a+c), (2)

if (a+c) > 0 otherwise Recall = 1

$$Accuracy = \frac{a+d}{a+b+c+d} \quad (3)$$

$$F-Measure = \frac{2*\Pr ecision*\mathrm{Re}call}{\Pr ecision+\mathrm{Re}call} \quad (4)$$

TABLE 6

CONFUSION MATRIX

| | *Correct Class $C_+$* | *Correct Class $C_-$* |
|---|---|---|
| Assigned Class $C_+$ | a | b |
| Assigned Class $C_-$ | c | d |

## 7. DOCUMENT PRE-PROCESSING

Not all the words in the documents are important, so they may degrade the classifier's performance. In addition, representing small set of documents that may have hundreds of different words using bag-of words approach will generate a huge feature space and thus will increase the processing time. To solve these problems, approaches to reduce the feature space dimension are needed. We used three approaches in the same sequence:

1. As a result of consulting an expert in the domain field, we removed unhelpful sentences from the documents such as "Informed consent was obtained with the benefits, risks and alternatives for the procedure explained", which is found in all reports;
2. We have removed stop words from all data sets using stop-lists containing common words such as "the", "a", "an";
3. We stemmed the words using Porter's suffix-stripping algorithm [3]. Words are considered the same if they share the same stem.

After performing above feature reduction approaches, the total number of features remained in the train set is 151337 features and 5167 distinct features remind. On the other hand, in the test set, the total number of features remained is 50041 features and 5167 distinct features.

## 8. METHODOLOGY SETUP

Once again, the primary concern of ours is to comprehensively evaluate different methods of partially supervised classification using real-world medical data. It will be possible to test the claim that his method is effective and computationally efficient [2] using a challenging medical problem. The combination of different methods used in step 1 (spy, NB and ROC) and step 2 (SVM and EM) will produce six techniques (classifiers) when we used one method for step1 and one method for step 2, for example, Roc-SVM technique means using Rocchio in step one and SVM in step two and so on. We divided our set of experiments into three phases each of which focuses on specific goals.

Phase I: The main goal of phase one is to evaluate the classification accuracy obtained by the six techniques under investigation for the problem of classifying the medical reports into successful and unsuccessful classes (data set way 1). In order to achieve this, the following sub-goals will be investigated:

Exp1.1- Investigating the effects of number of training samples accuracy (60, 120, 180, 240, 300 and 400) on the performance of the six classifiers. In traditional classification approach, the more training samples you have the more accurate the classifier is. So this set of experiments will show if this is the same case in partially supervised classification technique using medical data. In addition, more results mean more confidence about the behaviour of the techniques.

Exp1.2- Investigating the effects of six different choices for the number of training samples that yielded the best classification accuracy for all or most of the techniques. This set of experiments is depending on the output of (Exp.1.1). For example, if the number of training samples that produce best results is 180 samples, six different selections of 180 training samples will be used in order to have more confidence about the sample size we use.

Phase II: The top three techniques that produced best classification performance will be used in this phase for further extensive evaluation for different classification problems to find the best technique that perform better than the other two. This phase will use the second way of the data set which is the diagnosis. The diagnosises that will be used were described in the data set section. So the main goal of this phase is to compare the classification performance obtained by the top three techniques using different training samples (60, 120, 180, 240, 300 and 400) for four classification problems.

Phase III: Investigating the effect of using reduced set of features on the classification performance is the main objective of this phase. Two feature reduction/selection methods will be evaluated, term-frequency and term frequency-inverse document frequency (tf-idf).

## 9. RESULTS AND ANALYSIS

The results will be shown and analysed in the same sequence described in the methodology setup section.

### 9.1. Phase I

In this phase there are two main groups of experiments that will be conducted. Group 1 investigates the effects of using different number of training samples. Group 2 focuses on investigating the use of different selection of the same number of training samples.

1) Experiments (Group 1)

Table 7 bellow shows the F-Measure (FM) and accuracy (Acc.) results of SVM classifier as step two and ROC, NB and Spy methods for step one to identify reliable negative documents. The results obtained by using different number of training samples. The positive class recall and precision result were omitted due to space limitation. There are many observations could be made by analysing table 7. The very clear observation is that the more training samples used the better results obtained for both F-measure and accuracy. So these combinations behave similarly to the traditional classification approaches. Also it is very clear to note that there is a very significant improvement in the classification performance in term of F-measure and accuracy when the number of sample 120 is used for all techniques. For example, the F-measure and accuracy results obtained by Spy-SVM are improved by 4.8% in term of accuracy and by 29% in term of F-measure. For ROC-SVM and NB-SVM the accuracy improved by 2.5% and 3.5% respectively, and the F-measure improved by 13.2% and 30.8% respectively. The second significant improvement is obtained by NB-SVM using 180 samples, in this case, the accuracy improved by 4% and F-measure improved by 23.9%. The improvements in the classification performance obtained by the other training samples for all techniques are not significant when it compared to those two improvements. Comparing the classification performance obtained by ROC, NB and Spy using SVM method for step two we found ROC achieved the best results in term of accuracy and F-measure regardless the number of training samples used followed by Spy.

TABLE 7

F-MEASURE AND ACCURACY RESULTS OBTAINED BY SVM-BASED TECHNIQUES FOR DIFFERENT TRAINING SAMPLES

| Samples | ROC-SVM | | NB-SVM | | SPY-SVM | |
|---|---|---|---|---|---|---|
| | **FM %** | **Acc.%** | **FM %** | **Acc.%** | **FM %** | **Acc.%** |
| 60 | 68.84 | 92.95 | 24.63 | 87.45 | 47.66 | 89.91 |
| 120 | 82.05 | 95.41 | 55.42 | 90.89 | 76.70 | 94.67 |
| 180 | 83.65 | 95.73 | 56.00 | 90.98 | 83.54 | 95.73 |
| 240 | 88.36 | 96.80 | 79.87 | 95.00 | 84.18 | 95.90 |
| 300 | 89.68 | 97.13 | 85.00 | 96.06 | 86.07 | 96.31 |
| 400 | 90.29 | 97.21 | 89.68 | 97.13 | 90. | 97.21 |

On the other hand, F-Measure (FM) and accuracy (Acc.) results of EM classifier for step two and ROC, NB and Spy methods for step one to identify reliable negative documents for different number of training samples are shown in table 8. The combinations that use EM behave differently that those use SVM based on the classification performance. Using EM, for different number of training samples improve the performance to certain number of samples and then start decreasing when more training samples are used. This observation is true for the three techniques used EM. The main explanation of this observation is that in the medical report are similar to some extent. It contains mainly of number of organ and diagnosis names. The main difference is the medical staff description. So the classifier will be built based on the strong features the make maximum discrimination between the classes. That means, when the number of training sample sizes increase more weak features (weak feature means the features that found in positive and negative classes [31]) will be considered by the classifier and thus it leads to decrease the classification performance. In Table 8, we note that the best number of training samples for S-EM and ROC-EM that yielded the best classification performance is 120 samples in term of F-measure and accuracy. For this number of samples (120), the accuracy and F-measure obtained by S-EM is 95.98% and 85.88% respectively, and ROC-EM obtained 95.89% in term of accuracy and 85.96 in term of F-Measure. The 120 number of samples obtained second best result in term of accuracy and F-measure for NB-EM, whereas the best results for this technique obtained by using 180 samples based on accuracy (96.23%) and F-measure (86.93%) results. This is the same number of training samples that improve the classification performance significantly for NB-SVM technique when the accuracy jumped from 90.98% to 95.00% and the F-measure from 56% to 79.87%.

TABLE 8

F-MEASURE AND ACCURACY RESULTS OBTAINED BY EM-BASE D TECHNIQUES FOR DIFFERENT TRAINING SAMPLES

| Samples | ROC-EM | | NB-EM | | S-EM | |
|---|---|---|---|---|---|---|
| | **FM %** | **Acc.%** | **FM %** | **Acc.%** | **FM %** | **Acc.%** |
| 60 | 85.03 | 95.89 | 70.83 | 93.11 | 78.07 | 94.42 |
| 120 | **85.96** | **95.89** | 84.69 | 95.82 | **85.88** | **95.98** |
| 180 | 85.33 | 95.57 | **86.93** | **96.23** | 85.87 | 95.73 |
| 240 | 84.66 | 95.24 | 86.10 | 95.81 | 85.64 | 95.65 |
| 300 | 82.56 | 94.42 | 83.68 | 94.91 | 83.25 | 94.75 |
| 400 | 79.41 | 93.11 | 82.99 | 94.59 | 82.35 | 94.34 |

Comparing the classification performance obtained by ROC, NB and Spy using EM method for step two we found there is no method outperform the other two methods for all number of training samples. ROC outperforms the others when the number of training samples is 60. Spy obtained the best than the others when the number of samples is 120. NB achieved better results than the other two methods when the number of samples is greater than 120 samples.

As a conclusion of this set of experiment, all techniques use SVM behave similarly to the traditional classification techniques. Since we are investigating the effectiveness of different techniques using the small number of training samples, the SVM initially is not recommended.

2) Experiments (Group 2)

The first group of experiments that focused on evaluating the classification performance using different number of training samples, we could conclude that the best number of training samples that leads to better classification performance is 120 samples. Because it is yielded the best results for S-EM and ROC-EM, and the second of the best for NB-EM. In addition, the 120 samples improved the classification performance dramatically for all techniques used SVM, whereas the other number of samples greater than 120 improved the classification performance slightly as we observed from Tables 7 and 8.

Based on this observation, we used six different selections of 120 training samples for our new set of experiments for further and comprehensive evaluation of the six techniques. All techniques applied six times using one different sample of the six every time. Examining the techniques using different samples will proof whether the classifiers obtain consistent performance, moreover, show the impact of the training samples on the classification accuracy.

TABLE 9

AVERAGED VALUES AND STANDARD DEVIATION RESULTS FOR F-MEASURE AND ACCURACY OBTAINED BY ALL TECHNIQUES

| Techniques | FM Avg. | FM. S.D. | Acc. Avg. | Acc. S.D. |
|---|---|---|---|---|
| Roc-SVM | 81.035 | 2.164 | 95.177 | 0.457 |
| NB-SVM | 65.510 | 6.928 | 92.412 | 1.060 |
| Spy-SVM | 75.360 | 5.149 | 94.162 | 0.972 |
| ROC-EM | 84.325 | 1.285 | 95.273 | 0.535 |
| NB-EM | 82.570 | 2.445 | 95.210 | 0.544 |
| S-EM | 85.350 | 0.502 | 95.653 | 0.213 |

Due to the space limitation, the detailed results of this set of experiments were omitted even for F-measure and accuracy. Instead, we summarized the results using the simple average. We averaged the results of both F-measure and accuracy by the summing the six results of each of them for every technique and then we divided the total by 6. As the result, we got the averaged value of F-measure (FM Avg.) and the averaged value of accuracy (Acc. Avg.). The averaged values will show how well each technique perform, the greater averaged value the better result is. In addition to the averaged values, we used another statistical method namely standard deviation to measure the consistency of the results for each technique. We used the standard deviation because it has proven to be an extremely useful to measure how spread out the values in a data set are. More precisely, it is a measure of the average distance of the data values from their mean. If the values are all close to the mean, then the standard deviation will be low (closer to zero). If many values are very different from the mean, then the standard deviation is high (further from zero). If all the data values are equal, then the standard deviation will be zero [33]. We computed the standard deviation for both F-measure and accuracy. Table 9 shows the summary of the results containing the averaged values for F-measure and accuracy and the standard deviation for F-measure (FM S.D.) and accuracy (Acc. S.D.).

By analysing Table 9, there are many clear observations could be made. First, evaluating the techniques that use SVM with those uses EM, we find that EM techniques significantly outperform SVM techniques regardless the method used in step 1 based on averaged values of F-measure and accuracy. Moreover, EM techniques produce consistent results more than SVM techniques. For example, the difference between the best averaged accuracy and F-measure values and the worst results obtained by EM techniques are 0.4% and 2.8% respectively, and for SVM techniques are 2.8% and 15.5% respectively. For more evidence of consistency, we refer to the standard deviation results. Still EM techniques produce very consistent result based on Acc. S.D. and reasonable consistency results based on FM S.D. On the other hand, SVM techniques produce extremely inconsistency results in term of FM S.D. specially NB-SVM and Spy-SVM, the same observation is true in term of Acc. S.D. for the same two techniques.

TABLE 10

F-MEASURE AND ACCURACY RESULTS OBTAINED BY EM-BASE D TECHNIQUES FOR DIFFERENT TRAINING SAMPLES FOR CLASSIFICATION OF DIVERTICULOSIS DIAGNOSIS

| Samples | ROC-EM | | NB-EM | | S-EM | |
|---|---|---|---|---|---|---|
| | **FM %** | **Acc.%** | **FM %** | **Acc.%** | **FM %** | **Acc.%** |
| 60 | 55.56 | 88.20 | 29.66 | 84.84 | 36.44 | 39.67 |
| 120 | 69.57 | 90.82 | 45.58 | 86.89 | 66.84 | 89.51 |
| 180 | **78.94** | **92.21** | 53.63 | 87.95 | **70.53** | **89.59** |
| 240 | 76.64 | 91.80 | 61.45 | 89.10 | 60.45 | 81.23 |
| 300 | 76.13 | 90.90 | 65.22 | 89.51 | 59.80 | 80.49 |
| 400 | 73.92 | 89.59 | 65.81 | 89.10 | 55.31 | 75.49 |

Evaluating individual techniques, in term of Acc. Avg. and FM Avg. the best results obtained by S-EM followed by ROC-EM then NB-EM. For SVM techniques, still NB-SVM and Spy-SVM are the worst bad on the same measures. Based on the consistency of the obtained results, still S-EM outperforms all other techniques in term of FM S.D. and Acc. S.D. followed by ROC-EM. Although ROC-SVM produces more consistent result than NB-EM, still NB-EM is better because ROC-SVM is more consistent for worst FM S.D. and Acc. S.D. results than NB-EM. All these observations could be seen visually in Figure 5.

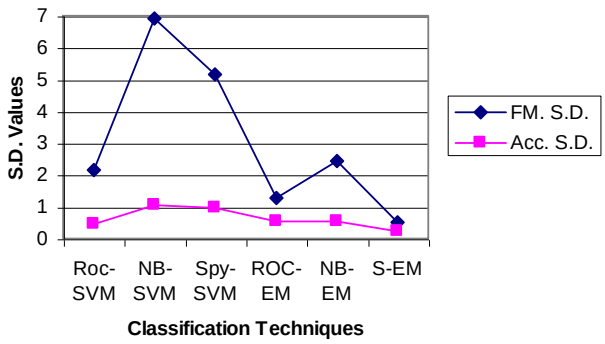


Figure 5. FM and Acc. Standard Deviation Results for all Techniques

As a conclusion of this phase, we found that the combination that uses EM in step two is always produces better results than those uses SVM using small set of training samples. In addition, EM techniques produce more consistent results as it explained above. So the top three techniques that will be investigated extensively in phase two are ROC-EM, NB-EM and S-EM.

TABLE 11

F-MEASURE AND ACCURACY RESULTS OBTAINED BY EM-BASE D TECHNIQUES FOR DIFFERENT TRAINING SAMPLES FOR CLASSIFICATION OF POLYPS DIAGNOSIS

| Samples | ROC-EM | | NB-EM | | S-EM | |
|---|---|---|---|---|---|---|
| | **FM %** | **Acc.%** | **FM %** | **Acc.%** | **FM %** | **Acc.%** |
| 60 | 83.15 | 91.07 | 21.80 | 72.95 | 83.92 | 90.98 |
| 120 | 89.34 | 93.93 | 43.48 | 77.62 | 82.15 | 90.49 |
| 180 | 91.82 | 95.16 | 56.38 | 81.23 | 85.50 | 90.57 |
| 240 | 92.54 | 95.49 | 65.01 | 83.85 | 88.26 | 92.54 |
| 300 | 93.15 | 95.82 | 70.61 | 85.74 | 88.98 | 93.03 |
| 400 | **93.39** | **95.90** | 76.25 | 87.95 | **91.29** | **94.51** |
| 500 | 92.79 | 95.49 | 82.50 | 90.57 | 91.00 | 94.26 |
| 600 | 91.86 | 94.92 | 84.23 | 91.31 | 89.23 | 93.03 |

TABLE 12

F-MEASURE AND ACCURACY RESULTS OBTAINED BY EM-BASE D TECHNIQUES FOR DIFFERENT TRAINING SAMPLES FOR CLASSIFICATION OF SESSILE POLYPS DIAGNOSIS

| Samples | ROC-EM | | NB-EM | | S-EM | |
|---|---|---|---|---|---|---|
| | **FM %** | **Acc.%** | **FM %** | **Acc.%** | **FM %** | **Acc.%** |
| 60 | 78.21 | 90.41 | 33.92 | 81.48 | 78.21 | 90.41 |
| 120 | 83.22 | 91.64 | 59.19 | 85.98 | 83.22 | 91.64 |
| 180 | **84.35** | **91.97** | 66.08 | 87.30 | **84.35** | **91.97** |
| 240 | 83.88 | 91.56 | 71.87 | 88.77 | 83.88 | 91.56 |
| 300 | 83.26 | 91.07 | 73.45 | 89.10 | 83.26 | 91.07 |
| 400 | 83.61 | 91.23 | 78.15 | 90.33 | 79.08 | 88.03 |

## 9.2 Phase II

The main input of this phase is the three techniques those produced better classification performance than the other three. The output is the recommendation of the best technique that obtain better results than the other two. Although the three techniques under investigation in this phase use the same data set, but they will be applied for different classification problems (diagnosis). These classification problems are a result of looking at or dealing with the same data set from different views (data set way 2) as we described in the dataset section. Four classification problems will be considered in this phase according to four diagnosis “Diverticulosis”, “Polyps”, “Sessile Polyps” and “Ulcerative Colitis”. Each one of these diagnosis is a classification problem. The results of applying the three techniques for four diagnosis classifications of Diverticulosis, Polyps, Sessile Polyps and Ulcerative Colitis using different number of training samples are shown in Tables 10, 11, 12, 13 respectively. We used different number of training samples because we are dealing with new classification problems and thus we don’t know which number of training samples produces the best results, in addition, it will give us more confidence about the final recommendation. Note that the number of training samples is not the same for all diagnosis classification, the reason is the maximum number of positive documents for each diagnosis are different (refer to Tables 2-5). Due to space limitation only F-measure and accuracy results will be shown.

TABLE 13

F-MEASURE AND ACCURACY RESULTS OBTAINED BY EM-BASE D TECHNIQUES FOR DIFFERENT TRAINING SAMPLES FOR CLASSIFICATION OF ULCERATIVE COLITICS DIAGNOSIS

| Samples | ROC-EM | | NB-EM | | S-EM | |
|---|---|---|---|---|---|---|
| | **FM %** | **Acc.%** | **FM %** | **Acc.%** | **FM %** | **Acc.%** |
| 15 | 20 | 93.44 | 0.41 | 93.36 | 39.58 | 90.49 |
| 30 | 56.41 | 94.03 | 13.33 | 93.61 | 30.51 | 93.28 |
| 60 | **61.38** | **94.12** | 43.08 | 93.93 | **61.62** | **94.18** |
| 90 | 61.03 | 93.20 | 58.23 | 94.59 | 58.88 | 93.36 |
| 120 | 58.82 | 92.54 | 60.71 | 94.59 | 58.94 | 93.03 |

Comparing the results of four diagnosises classification problem (shown in Tables 10 to 13) with the results of successful/failed procedures (shown in Table 8) we noticed the following (1) in Table 8, only 120 samples are needed to reach the best performance according to ROC-EM and S-EM many. The same two techniques needed at least 180 samples to reach the best performance. One possible reason is there are more positive documents in Table 8 (656) than the positive documents of the Diverticulosis (910), Polyps (1553) and Sessile Polyps (1190) (refer to Tables 1, 2, 3, 4 and 5). But this is not the right reason, because even if more training samples are needed, the best accuracy should be as good as in Table 8 or at least -/+ 1% but not 89.50% in term of accuracy and 61.4% in term of F-measure in some cases. More evidence, the 120 samples in Ulcreative Colitics should yielded extremely better results than in Table 8 if the reason is the number of positive documents (refer to Tables 1 and 5). In our opinion, we believe the right reason is the number of strong features that discriminate very well between the positive and the negative classes. That means, there are considerable number of strong features to distinguish between the two classes in the case of successful and failed classification problem in phase one. (2) Referring to Tables 10 to 13, we note that NB-EM in all these experiments is behaving similarly to traditional classification approaches, based on the relationship between the number of sample size and the classification performance. The main explanation of this is NB is failed to identify a good set of reliable negative documents using small set of training samples. Thus, the classifier in step two will have few strong features and many weak features to build the classification model. So that means more samples include more strong features and thus leads to better classification performance. This also explain the following (a) in Table 8 NB-EM needed 180 samples to obtain the best results whereas the other two needed only 120, (b) there are considerable amount of strong features in the case of successful and failed classification problem, so when more samples used more weak features will be included in the classification model. This is opposite situation in the case of classifying diagnosis, there is considerable amount of weak features, and thus more samples will include strong features. All above observation could be noticed clearly in the exceptional case of diagnosis which is the Polyps in Table 11.

TABLE 14

F-MEASURE AND ACCURACY RESULTS OBTAINED BY ROC-EM AND S-EM FOR SUCCESSFUL/FAILED CLASSIFICATION USING TOP 100, 200, 300, 400 AND 500 FEATURES

| Features | ROC-EM | | S-EM | |
|---|---|---|---|---|
| | **FM %** | **Acc.%** | **FM %** | **Acc.%** |
| All | 85.96 | 95.89 | 85.88 | 95.98 |
| Top 100 | 76.85 | 92.54 | 73.37 | 91.31 |
| Top 200 | 87.89 | 96.48 | 88.39 | 96.64 |
| Top 300 | 89.64 | 96.97 | 89.01 | 96.80 |
| Top 400 | **90.17** | **97.21** | **89.74** | **97.13** |
| Top 500 | 89.47 | 97.05 | 88.18 | 96.64 |

In this table, the number of training samples (400) that obtained the best classification performance according to ROC-EM and S-EM reflect the big number of positive documents (1553). In term on the considerable amount of strong features could be noticed by the classification accuracy obtained by ROC-EM (95.9%), for example, is very competitive to those obtained by ROC-EM and S-EM (95.89% and 95.89% respectively), and by the excellent classification performance in term of F-measure obtained by ROC-EM and S-EM (93.39 % and 91.29 % respectively) are much better that those obtained by the same techniques in Table 8 which are 85.90 and 85.88 respectively. (3) ROC-EM and S-EM are always behave exactly the same in term of the relationship between the number of training samples and the classification performance in term of F-measure and accuracy including the best results obtain by both techniques in term of F-measure and accuracy use the same number of training samples.

As a conclusion of this phase, we found that the technique NB-EM is beehives similarly to the traditional approach. In addition, the techniques ROC-EM and S-EM are always obtain very competitive classification performance based on F-

measure and accuracy in all the classification problems. Both techniques will be used in phase three to investigate the effects of using reduced set of features on the classification performance.

TABLE 15

THE NUMBER AND PERCENTAGES OF EMPTY DOCUMENT AND REMAINING FEATURES FOR DIFFERENT THRESHOLDS

| Selection Criteria | Empty Documents | | Remaining Features | |
|---|---|---|---|---|
| | # | % | # | % |
| tf-idf >= 0.2 | 7 | 0.19 | 24887 | 19.81 |
| tf-idf >= 0.15 | 1 | 0.03 | 39965 | 31.81 |
| tf-idf >= 0.125 | 1 | 0.03 | 48846 | 38.88 |
| tf-idf >= 0.1 | 0 | 0 | 58168 | 46.30 |
| tf-idf >= 0.08 | 0 | 0 | 87757 | 69.85 |

## 9.3. Phase III

The above two phases included the full set of features found in all documents in the data set. This phase focuses mainly on the effect of using reduced features on the classification performance. The final total number of unique features in the collection is 5,167. The frequencies of these features vary from the highest frequency 7111 to the lowest frequency of 1. Over 2400 of these features occurred only once. There are many feature-reduction/features-selection approaches such as mutual information and information gain could be used. In this set of experiments we investigated two very simple methods. First method reduces the features based on their frequencies (term-frequency). The second method reduces the features based on the term frequency-inverse document frequency (tf-idf) values. Using term-frequency method, only the $\gamma$ top features according to their frequency will be selected to build the classifier. The five values of $\gamma$ used are 100, 200, 300, 400 and 500. The tf-idf method, the tf-idf values ranging from 0.00001 to 1.0. In order to reduce the features, different tf-idf threshold values should be investigated to choose the threshold that is not result in big number of empty documents (the document that all its features will not be selected).

TABLE 16

F-MEASURE AND ACCURACY RESULTS OBTAINED BY ROC-EM AND S-EM FOR SUCCESSFUL/FAILED CLASSIFICATION USING DIFFERENT *TF-IDF* THRESHOLDS

| ***tf-idf th*** | ROC-EM | | S-EM | |
|---|---|---|---|---|
| | **FM %** | **Acc.%** | **FM %** | **Acc.%** |
| All | 90.17 | 97.21 | 89.74 | 97.13 |
| tf-idf >= 0.2 | 19.91 | 86.14 | 9.52 | 85.97 |
| tf-idf >= 0.15 | 34.48 | 87.53 | 20.31 | 87.12 |
| tf-idf >= 0.125 | 40.18 | 88.77 | 24.51 | 87.38 |
| tf-idf >= 0.1 | **47.62** | **89.18** | **28.17** | **87.46** |
| tf-idf >= 0.08 | 44.09 | 88.36 | 23.26 | 86.48 |

TABLE 17

F-MEASURE AND ACCURACY RESULTS OBTAINED BY ROC-EM FOR DIVERTICULOSIS AND POLYPS CLASSIFICATION USING TOP 100, 200, 300, 400 AND 500 FEATURES

| Features | Diverticulosis | | Polyps | |
|---|---|---|---|---|
| | **FM %** | **Acc.%** | **FM %** | **Acc.%** |
| All | 78.94 | 92.21 | 93.39 | 95.90 |
| Top 100 | 78.24 | 91.89 | 92.47 | 95.33 |
| Top 200 | 78.94 | 92.21 | **94.00** | **96.23** |
| Top 300 | **80.45** | **92.79** | 93.86 | 96.15 |
| Top 400 | 80.18 | 92.79 | 93.98 | 96.23 |
| Top 500 | 78.72 | 92.38 | 93.33 | 95.82 |

This phase will proof the following: (1) if reducing the features will improve the classification performance of partially supervised approach using medical documents (2) which feature reduction method is perform better than the other. To achieve this, first, we used the two techniques recommended by phase two (ROC-EM and S-EM) for the successful/failed classification problem using term frequency method first then tf-idf. The best feature selection method will be used by ROC-EM for the diagnosis classification problem namely Diverticula and Polyps. Using ROC-EM as a classification technique and Diverticula and Polyps as diagnosis classification problems just to have more confidence that reducing the features improving the classification performance or not, and not based on any recommendations.

Table 14 shows the results of the classification performance obtained by the two techniques using different 5 values of $\gamma$. In the same table, the first row shows the best F-measures and accuracy results obtained by the same techniques using all features to facilitate the comparison process. Due to space limitation we show only the F-measure and accuracy results. From this table, we could notice clearly that the reduced number of feature improve the classification performance significantly in term of F-measure and accuracy for both techniques. ROC-EM improves F-measure and accuracy by 4.2% and 1.32% respectively, and S-EM by 3.86% and 1.15% respectively. Figure 6 shows graphically the improvement on classification performance in term of F-Measure and accuracy using reduced set of features. Using the top 100 features is significantly degraded the classification performance. This may indicate that a set of 100 features is too small to produce and revise good probabilistic labels of the documents in U-RN when using EM method.

Table 15 shows the number and percentages of empty documents for different thresholds (th). Note that the total number of training documents is 3657 and the total number of features is 125630. We have check more thresholds, due to space limitation we showed some of them. Table 16 shows the results obtained by the two techniques using different values of tf-idf thresholds. The best results achieved for both techniques using threshold = 0.1. But the best results obtained in terms of F-measure and accuracy are extremely worst than using all features as shown in the first row of the same table. The results shown in Table 17 illustrate the classification performance of Diverticula and Polyps classification problems using five values of $\gamma$. This table emphasising the effectiveness of using reduced feature on the classification performance. The best number of top features is 200 for Polyps and 300 for Diverticula.

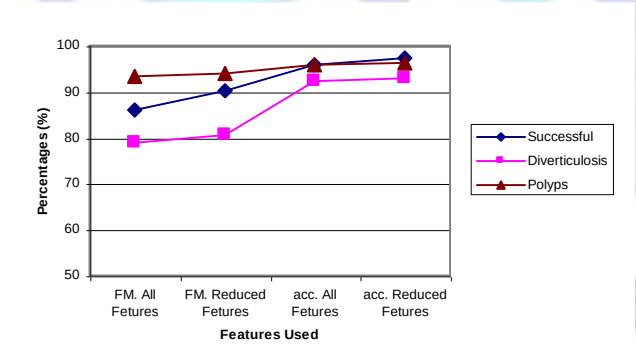


Figure 6. F-Measure and Accuracy Results obtained Using All and Reduced set of Features for three Classification Problems

As a conclusion of this phase, reducing the features based on their tf-idf values is decreasing the classification performance dramatically. Also using a very reduced set of features degrade the classification performance. Moreover, reducing the features based on their frequencies improve the classification performance significantly if the features are selected carefully. Improving the classification performance by using reduced set of features support our opinion in phase two when we rely the low classification performance obtained in phase two than in phase one to the quality of the features (weak or strong) and not to the number of positive document. Final conclusion, finding a sufficient set of features can improve performance while also increasing efficiency, but it may require some experimentation.

## 10. CONCLUSION

The objective of this paper is to comprehensively investigate the partially supervised classification approach on a real world problem, especially in the medical field. To achieve this, five classification problems were conducted to evaluate the classification performance using different methods within the two-step approach. Some techniques approved its efficiency of producing high classification performance using only a small set of labeled positive documents to operate. Our experimental results showed that the combination that uses EM in step two is always produces better results than those uses SVM using small set of training samples. In addition, EM techniques produce more consistent results. The partially

supervised classification techniques that use SVM in step two and NB-EM technique behave similarly to the traditional classification techniques, the more training samples the better result obtained. Since we are investigating the effectiveness of different techniques using the small number of training samples, the SVM initially is not recommended for classification of medical reports unless there is enough training samples. The combination that always obtain very competitive classification performance based on F-measure and accuracy in all the classification problems are ROC-EM and S-EM.

We experimentally showed that reducing the features based on their tf-idf values is decreasing the classification performance dramatically. And using a very reduced set of features degrade the classification performance as well. Moreover, reducing the features based on their frequencies improve the classification performance significantly while also increasing efficiency, but it may require some experimentation.

Our results are very competitive for this real world problem and could be used to automatically label and classify medical reports. We believe the method is widely applicable to other text classification problems in the medical domain that requires two-class or binary classification.

## REFERENCES


[1] B. Liu, Y. Dai, X. Li, W. S. Lee, and P. Yu, “Building Text Classifiers Using Positive and Unlabeled Examples”. Proceedings of the Third IEEE International Conference on Data Mining (ICDM-03), Melbourne, Florida, 2003.

[2] B. Liu, W. S. Lee, P. S. Yu, and X. Li, “Partially Supervised Classification of Text Documents”. Proceedings of the Nineteenth International Conference on Machine Learning (ICML-2002), Sydney, Australia. 2002.

[3] M. F. Porter, "An algorithm for suffix stripping", Program; automated library and information systems, 14(3), 130-137, 1980.

[4] H. Benbrahim, and M. A. Barmer, “Neighborhood Exploitation in Hypertext Categorization”. In Research and Development in Intelligent Systems XXI. Springer-Verlag, 2005.

[5] B. D. Aronow, and F. Feng, “Ad-Hoc Classification of Electronic Clinical Documents”. D-Lib Magazine. ISSN 1082-9873. 1997.

[6] C. J. Bowles, R Leicester, C. Romaya, E Swarbrick, C. B. Williams, and O. Epstein, “A Prospective Study of Colonoscopy Practice in the UK today: are we Adequately Prepared for national colorectal Cancer Screening Tomorrow?” International Journal of Gastroenterology and Hepatology, 2003.

[7] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell, “Learning to Classify Text from Labeled and Unlabeled documents”. AAAI-98. pp 792-799. AAAI Press. Menlo Park, US. 1998.

[8] Y. Yang, and X. Liu, “Are-examination of Text Categorization Methods”, Special Interest Group of Information Retrieval (SIGIR), 1999.

[9] D. D. Lewis, “Representation and Learning in Information Retrieval”, PhD Thesis, Department of Computer and Information Science, University of Massachusetts, 1992.

[10] Jiawei Han, Micheline Kamber, Jian Pei (2012), Data Mining Concepts and Techniques: Elsevier

[11] J. Rocchio, “Relevant Feedback in Information Retrieval, The smart retrieval system-experiments in automatic document processing”. Englewood Cliffs, NJ, 1971

[12] A. McCallum, and K. Nigam, “A Comparison of Event Models for Naïve Bayes Text Classification”. In AAAI-98 Workshop on Learning for Text Categorization, 1998.

[13] L. Xiaoli, and B. Liu, “Learning to classify text using positive and unlabeled data”. Proceedings of Eighteenth International Joint Conference on Artificial Intelligence (IJCAI-03), Acapulco, Mexico. 2003.

[14] D. D. Lewis, “Evaluating Text Categorization”. Proceedings of the Speechand Natural Language Workshop Asilomar, Morgan Kaufmann, pp 312-318. 1991.

[15] B. S. Raghavendra and Ajit S. Bopardikar, “Identification of CpG Islands in DNA Sequences using Supervised Classification”. Proceeding of IEEE International Conference on Bioinformatics and Biomedicine Workshops, 2011.

[16] A. Dempster, N. M. Laird, and D. Rubin, “Maximum Likelihood from Incomplete Data via EM Algorithm”. Journal of the Royal Statistical Society, 1997

[17] D. Lewis, and M. Ringuette, “A Comparison of Two Learning Algorithms for Text Categorization”. 3rd annual symposium on document analysis and information retrieval, pp. 81-93, 1994.

[18] T. Joachim, “Making Large Scale SVM Learning Practical”. Advances in Kernel Methods - Support Vector Learning, 1999.

[19] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell, “Text Classification from Labeled and Unlabeled Documents Using EM”. Machine Learning, pp. 103-134, 2000.

[20] G. Salton, and M. McGill, “Introduction to Modern Information Retrieval”. McGraw-Hill. 1983.

[21] Z. Hua-Jun, W. Xuan-Hui, Z. Chen, H. Lu, and M. Wei-Ying, "CBC: Clustering Based Text Classification Requiring Minimal Labeled Data". Third IEEE International Conference on Data Mining (ICDM'03), pp. 443-450. 2003.

[22] A. Blum, and T. Mitchell, “Combining Labelled and Unlabeled Data with Co-Training”. In Proceedings of the 11th Annual Conference on Computational Learning Theory. pp. 92-100. 1998.

[23] T. Joachims, “Transductive Inference For Text Classification Using Support Vector Machines”. In Proceedings of 16th International Conference on Machine Learning. pp. 200-209. 1999.

[24] R. Bhavani, F. Herman, and A. Kowalczyk, “Combining Clustering and Co-Training to Enhance Text Classification Using Unlabeled Data”. In Proceedings of the SIGKDD International Conference on Knowledge Discovery and Data Mining. 2002.

[25] K. Nigam, and R. Ghani, “Analyzing the Effectiveness and Applicability of Co-Training”. In Proceedings of 9th International Conference on Information and Knowledge Management. 2002.

[26] R. Ghani, “Combining Labelled and Unlabeled Data for Multi-Class Text Categorization”. In Proceedings of the International Conference on Machine Learning (ICML). 2002.

[27] S. Goldman, and Y. Zhou, “Enhanced Supervised Learning with Unlabeled Data”. In Proceedings of the International Conference on Machine Learning (ICML). 2000.

[28] S. Basu, A. Banerjee, and R. Mooney, “Semi-Supervised Clustering by Seeding”. In Proceedings of the International Conference on Machine Learning (ICML), 2002.

[29] J. Bockhorst, and M. Craven, “Exploiting Relations among Concepts to Acquire Weakly Labelled Training Data”. In Proceedings of the International Conference on Machine Learning (ICML), 2002.

[30] I. Muslea, S. Minton, and C. Knoblock, “Active + Semi-Supervised Learning = Robust Multi-View Learning”. In Proceedings of the International Conference on Machine Learning (ICML), 2002.

[31] H. Yu, J. Han, and K. Chang, “PEBL: Positive Example Based Learning for Web Page Classification Using SVM”. KDD-02, 2002.

[32] B. Scholkopf, J. Platt, J. Shawe, and A. Smola, Williamson R., “Estimating the Support of a High-Dimensional Distribution“. Technical Report MSR-TR-99-87, Microsoft Research, 1999. Available at http://www.cs.cmu.edu/~aarnold/ids/postal.pdf

[33] B. S. Everitt, “Cambridge Dictionary of Statistic in Medical Sciences”. Cambridge University Press. 1995.

[34] F. H. Saad, B. de la Iglesia, G. D. Bell, “Comparison of Documents Classification Techniques to Classify Medical Reports”. Advances in Knowledge Discovery and Data Mining, Lecture Notes in Computer Science Volume 3918, 2006, pp 285-2912008

[35] F. Denis, “PAC Learning from Positive Statistical Quires”, ALT, pp 112-126. 1998