A Feature Learning based Technique to Classify Medline Disease Abstracts

Hisham Al-Mubaid
Dept. of Computer Science
University of Houston – Clear Lake
Houston, USA
hisham@uhcl.edu

Andrew Nash
Dept. of Computer Science
University of Houston – Clear Lake
Houston, TX
USA

***Abstract*—In bioinformatics, the classification of research documents in relation to their subject matter (specifically, disease types) is a very important task with many applications in the field. In terms of text classification, the total number of available disease documents is limited with *Medline* being the main source, containing a mostly abstract-only collection of research papers. However, abstracts are an important tool in gauging the specific disease in question, and giving a quick summary of the research. We introduce an effective technique based on feature learning for inducing feature weights appropriate for text classification with small dataset (in training and testing). We applied and evaluated the proposed technique in text classification task with medical abstracts representing disease texts. The conducted experiments with small data sets ranging from 40 to 500 documents (abstracts composed of around 300 words) produced excellent and promising results with respect to classification accuracy and AUC. In essence, the proposed technique gives leverage to the class distribution over each attribute rather than the attribute distribution over the two classes. Our technique consistently showed higher performance than conventional methods, and continued to improve with a decrease in the number of documents across all disease abstract classification experiments.**

Keywords—disease document classification, feature learning.

I. Introduction

he text classification task in bioinformatics is highly important and can significantly impact many problems in this domain such as medical information retrieval, information indexing, gene-disease association, the understanding of gene functions, and biomedical knowledge discovery [12, 20, 28, 36]. In this sort of text classification task, we would like to assign new class labels for unclassified text documents based on a set of pre-defined class labels. Text classification is one of the highly active research areas in several fields including bioinformatics [1, 2, 12, 15, 20, 25, 26, 27, 28]. In this paper, we introduce a new technique for biomedical text classification of disease abstracts based on calibrating the contribution of words and terms in the text corpora. The classification task of text documents modeled by *bag-of-word* can be casted as an induction problem in a very high dimensional space. Our technique in this paper depends on the probability of the class conditioned on the attributes. Basically, we employ the difference in the probability of a class given the presence or absence of an attribute, applied independently for each attribute in the corpus. This supervised scheme shall enhance the calibration of each attribute, *i.e., term*, according to its contribution in each class so as to improve the prediction performance.

Each attribute (term) is weighted/evaluated based on its presence in a document without giving attention to the number of times that the term occurred in the document. In biomedical text classification, this is an appropriate and effective approach as the inclusion of a term in a given document is significant regardless of the frequency of which that term appears.

In a disease-focused medical text mining task, there are a limited number of available documents (*i.e.,* disease abstracts), and therefore a relatively small number of terms (*i.e.,* attributes). Because of this, our technique leverages the existence of each term in a given abstract, transformed with respect to the class probabilities given the existence and absence of the term, normalized by the overall log probability of the term.

We have conducted a large number of experiments to evaluate our technique using several experimental settings and many datasets of disease abstracts from *Medline*. The results of our proposed technique are promising and proved effective for inducing fairly strong text classifiers in this domain. Our technique outperformed the baseline method in almost every class pair across all experiments. The strength of our technique becomes even more evident as the number of documents decreases {*note: we used the words 'feature' and 'attribute' interchangeably to refer the same*}.

II. Background and Related Work

The biomedical texts produced from medical research and clinical trials are growing at higher rates year after year [3, 7, 8]. Therefore, it is becoming more important to catalogue and organize this biomedical research so that relevant information can be retrieved more easily. In [18], Donaldson et al. (2015)

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

gathered volumes of *Medline* abstracts to categorize the information in the texts on protein–protein interactions. They then curated and added the information to their BIND database [18]. Using text classification techniques, they were able to reduce the number of abstracts that the curators needed to read by about two-thirds.

A *Kullback-Leibler* based Probabilistic Latent Categorizer (PLC) method was proposed in [19]. They re-ranked documents for the purpose of curating information in the *Swiss-Prot* database. In this case, the results improved precision by 25 – 45% over the *PubMed* ranking methods [19]. In [31], Liu et al. (2004) proposed a classification method to aid in the search of gene and protein interaction information that already exists in the figures of full text research documents [31].

Several attribute-based methods have been proposed for attribute encoding and representation such as *TFIDF* term frequency-inverse document frequency *IDF*, category concept, as well as other concepts [1, 2, 12, 15, 20, 25, 26, 27]. For example, the *IDF-based* methods evaluate a given attribute based on an inverse relationship with the number of documents containing that attribute (term). When the number of documents containing a specific attribute increases, the capability of such an attribute in effectively discriminating classes decreases. In the *Information Retrieval* domain, this is valid, but it needs some modification for use with text classification. Specifically, when the number of documents containing some attribute t_k increases, and most of those documents belong to class C_j, then the attribute t_k should be considered a powerful factor for discriminating class C_j from the other class. TFIDF is the most successful and most widely used attribute weighting and encoding method in text classification [1, 2, 12, 15, 20, 25, 26, 27].

In dimensionality reduction and feature selection, one removes all unnecessary and redundant attributes in the vector space [1, 2, 4]. In text classification, feature selection and reduction techniques have been extensively used in many bioinformatics tasks such as the classification of microarray data, gene selection, biomedical document clustering, gene-protein function prediction, gene-protein name disambiguation, and biomedical term disambiguation [7, 8, 12]. Furthermore, many applications in *Natural Language Processing* have improved significantly from the reduction and selection techniques, including text categorization [7, 8].

III. New Feature Learning Technique

We introduce a new technique based on feature learning and inducing effective representation of data for improved learning and classification performance. We applied the method for the classification of medical texts from *Medline* [16] representing disease abstracts. In a document classification task, given a target document and a set of predefined class labels, we would like to assign a class label for the document. The key contribution of this paper is in the way we induce the encodings of the terms (attributes) to improve the classification performance when the size of the corpus (number of documents) is not large. Our technique can encode document term attributes effectively for accurate learning and prediction. Each *term* in this work is basically an attribute in the classification task. We utilize the class probability distribution in our technique, rather than attribute distribution. The class distribution is conditioned on the presence versus absence of the attribute, and for all attributes. Then the class probability distribution is computed by taking the difference in class probability *with* versus *without* the attribute. That is, we investigate and utilize the difference between class probability in existence of attribute t_i $P(C|t_i)$ *versus* in absence $P(C|\sim t_i)$ for each attribute t_i. Each term in the text corpus is considered one attribute. The technique relies on vector space representation of documents [1, 2, 12, 15, 20, 25, 26, 27, 28]. Each term in the document corpus is a component in an *n*-dimensional vector where *n* is the total number of terms in the entire document corpus. Let the vector X_i represent document d_i, each term (attribute) t_j in d_i can be represented as a component in vector X_i.

Let T = $\{t_i\}_{i=1..n}$ be the set of all attributes in a given text dataset. Each attribute t_i represents a term (*a word*) from the text corpora after some text preprocessing operations, and *n* represents the total number of terms in the entire text collection. The preprocessing steps remove all stop-words, all non-word tokens (like numbers), and also process inflicted word forms through stemming. For a given attribute t_i in a 2-class classification $(C, \sim C)$ the weight/encoding of attribute t_i, called w_i, is computed based on the probability of class *C* conditioned on the presence and absence of the attribute as follows:

$$w_i = \frac{P(C|t_i) - P(C|\sim t_i)}{1 - \frac{logP(t_i)}{2}} \qquad (1)$$

and this can also be written as:

$$w_i = (\frac{P(t_i,\ C)}{P(t_i)} - \frac{P(\sim t_i,\ C)}{P(\sim t_i)})/(1 - \frac{logP(t_i)}{2}) \qquad (2)$$

where $P(C\,|t_i)$ is the probability of class *C* conditioned on attribute t_i; and $P(C\,|\sim t_i)$ is the probability of class C conditioned on absence of t_i; therefore, from equations (1) and (2), we have:

$$-1.0 \le w_i \le 1.0 \qquad (3)$$

Even though we consider the two-class case here, we can also account for the *k*-class case by inducing *k* classifiers during the training step, one for each class. In equation (3), we have $w_i \in [-1\ ..\ 1]$, where an attribute t_i tends to incline to class *C* (resp. to class ~*C*) as its encoding w_i approaches 1.0 (resp. –1.0). Let X_i be the numeric vector of document d_i such that:

$$X_i = \{x_{ij}\}_{i=1..m}^{j=1..n}$$

where x_{ij} is the value of *jth* attribute t_j in the vector X_i, *n* is the total number of attributes, and *m* is the total number of documents. Then, $P(C\,|\,t_j)$ can be depicted as:

$$P(C|t_j) = \frac{|\,X_i\colon X_i \in C\ and\ x_{ij} \neq 0|}{|\,X_i\colon\ x_{ij} \neq 0|} \qquad (3a)$$

Moreover, in equation (1), $P(t_i)$ is calculated as:

$$P(t_i) = \frac{|\,X_q\colon\ x_{qi} \neq 0|}{m} \qquad (3b)$$

where *m* is the total number of documents in the collection. We use the most famous and commonly used TFIDF (*i.e., term frequency-inverse document frequency*) method in the *baseline* method for all attributes where:

$$tfidf_{ij} = tf_{ij} * idf_j \qquad (4)$$

where tf_{ij} is the term frequency if term t_j in document d_i, and idf_j is the inverse document frequency of term t_j as follows: $idf_j = \log(N \,/\, df_j)$ and df_j (*document frequency*) is the number of documents containing the term t_j .Then, attribute encodings in the *baseline* will be:

$$x_{ij} = tfidf_{ij} \qquad (5)$$

In our proposed technique:

$$x_{ij} = w_j . tf_{ij} \qquad (6)$$

The baseline method, *tfidf*, is the most widely used and highly successful in the encoding and feature weighting task in text classification problem. This proposed technique is suitable for a document classification task where number of documents are small like disease abstracts and each term weight is calibrated accordingly as in equation (2) because it is mainly based on attribute appearance $\{P(C \mid t_i)\}$ versus absence $\{P(C|{\sim}t_i)\}$ normalized by the log of the attribute probability. Specifically, in the case of disease abstracts, the appearance of an attribute in a given text is more significant based on the how the class probability is distributed among abstracts that contain this word, and those that do not.

IV. EXPERIMENTS AND RESULTS

We have conducted text classification experiments to evaluate our technique using datasets of disease abstracts from the *Medline* database [16]; see Table 1. *Medline* is the world's largest and most comprehensive biomedical database with more than 25 million records, covering all life science fields [16, 19]. Some of these records include full text articles but mostly only contain the abstracts of biomedical research articles [16]. The work in this paper is focused on disease abstracts in the *Medline* database. We retrieved the disease abstracts using the *Entrez–PubMed* search and retrieval interface. We obtained disease names (other disease terms, other names) from the *Disease Ontology* (DO: http://disease-ontology.org/) and topic C (diseases) in *MeSH* (https://meshb.nlm.nih.gov/treeView) database.

We used the Bayesian method for training and classification [10, 11]. We also initially experimented with *SVM* [1, 5, 9], *J48* (*Decision tree*), and *Random Forest*, and found *Bayesian* giving the most stable performance overall (*explained later, in section V. Discussion, see Table 9 and Figure 6*). We examined the performance of our technique by comparing the classification accuracy, *F1* measure, and *AUC* against a baseline method, both using *10-fold cross-validation*.

<u>Disease abstract datasets</u>: The disease text datasets are summarized in Table 1 (for the first batch of experiments). Since specific diseases do not have many abstracts explicitly about the target disease, we tailored our search to retrieve and test a small number of documents ranging from 40 to 500 as shown in Table 1. For example, the 500 setting, *Setting-4* in Table 1, is tailored towards a disease that can easily retrieve at least 250 abstracts essentially about that disease.

For example, searching Medline for the last two years for disease: *T-cell acute lymphoblastic leukemia* (MeSH Unique ID: D054198; Disease Ontology Id: DOID:5602), we found only 436 abstracts in the past 2 years; also, we found 146 abstracts in the past five years for Mastocytoma (MeSH Unique ID: D034801; Disease Ontology DOID:3664; also known as *Mast cell Proliferative disease*, and *mast cell tumor*). We performed all the necessary pre-processing steps (including word stemming and stop-word removal [4, 29]).

<u>Evaluation metrics</u>: The classification *Accuracy* is the first metric we use and report and can be depicted as:

$$Accuracy = \frac{Correctly\ classified\ documents}{Total\ number\ of\ tested\ documents} \qquad (7)$$

We also report *Area Under ROC Curve* (AUC) in the evaluation. The ROC curve (receiver operating characteristics) is a curve of the true positive rate *TPR* (x-axis) against the false positive rate *FPR* (y-axis). In general, AUC is a reliable indicator of the performance of a classification system is commonly adopted for use in machine learning.

<u>Experiments and Results</u>: We conducted all the experiments using *10-fold cross-validation* with the *Bayesian* method as it produced the most stable performance over the varied number of documents (*from 40 to 500*); See Table 9 and Figure 6. In the first evaluation batch, we conducted 38 experiments comprising of 7400 total disease abstracts (Table 1). Then we conducted another batch of experiments with 3912 documents as shown in Table 7; and the last batch of experiments (Shown in Table 8) included 9 experiments to examine the performance of the proposed technique while varying (decreasing) number of documents. The conducted experiments and results are summarized in Tables 2 – 5. These results are also illustrated in Figures 1 – 3. The accuracy and AUC of *Setting–2* (100 documents) and *Setting–3* (200 documents) are depicted in Figure 3 and Figure 4 respectively. As shown in Tables 2 – 5 , the proposed technique produced higher accuracy and AUC results in almost all experiments, and in many times with significant improvement.

TABLE 1: The experimental settings of the Medline disease abstract datasets

Experimental Setting	Total Number of Abstracts	Percentage of Disease 1 Abstracts	Average Number of Attributes	Number of Experiments
Setting–1	40	~50%	430	10
Setting–2	100	~50%	920	10
Setting–3	200	~50%	3760	10
Setting–4	500	~50%	5350	8

TABLE 2: Performance results in terms of *Accuracy* and *AUC* of the proposed and the baseline with the first setting, *Setting-1*, that includes only 40 disease abstracts.

Experiment (Disease-1: Disease-2)	**Total Number of Attributes**	**Accuracy**		**AUC**	
		Baseline	**Proposed**	**Baseline**	**Proposed**
Alzheimer : Parkinson	414	0.725	0.875	0.752	0.907
Alzheimer : Parkinson #2	414	0.825	0.975	0.865	0.980
Alzheimer : Diabetes	466	0.814	0.850	0.822	0.868
Influenza : Measles	444	0.805	0.815	0.863	0.881
Asthma : Bronchiolitis	442	0.850	0.925	0.850	0.915
Autism : Asperger	442	0.800	0.875	0.808	0.870
Celiac : Crohns	397	0.850	0.900	0.879	0.895
Dyslexia : Dyspraxia	389	0.900	0.950	0.881	0.918
Graves : Goiter	406	0.750	0.875	0.798	0.865
Kidney Disease : Fatty Liver	505	0.875	0.950	0.920	0.965
Total Number of Documents: 40 (*20 Disease-1 + 20 Disease-2*)					

TABLE 3: Performance results using Setting-2 where each class contains 50 disease abstracts.

Experiment (Disease-1: Disease-2)	**Total Number of Attributes**	**Accuracy**		**AUC**	
		Baseline	***Proposed***	***Baseline***	***Proposed***
Heart Disease : Stroke	915	0.575	0.610	0.556	0.660
Asthma : Bronchiolitis	935	0.860	0.890	0.870	0.915
Autism : Asperger	843	0.670	0.730	0.705	0.780
Celiac : Crohns	987	0.870	0.890	0.944	0.958
Dyslexia : Dyspraxia	881	0.890	0.920	0.894	0.920
Fibromyalgia : Polymyalgia	924	0.880	0.880	0.927	0.940
Graves : Goiter	879	0.770	0.850	0.838	0.890
Hepatitis : AIDS	994	0.820	0.880	0.876	0.910
Kidney Disease vs. Fatty Liver	934	0.790	0.840	0.891	0.925
Obesity vs. Diabetes	900	0.740	0.790	0.748	0.795
Total Number of Documents: 100 (*50 Disease-1 + 50 Disease-2*)					

The second batch of evaluation experiments are shown in Table 7 and discussed in Section V; also discussed in the next section is the last batch of 9 disease abstract classification experiments. We eliminated any classification task (experiment) that results in less than 100 words or less than 25 distinct words after removing stop words . We noticed that diseases with low number of publications also tend to have relatively low number of genes associated with them. For example, disease '*CRC somatic Colorectal Cancer*' (OMIM #114500) is associated with 16 genes; whereas disease '*Feingold syndrome*' OMIM# 164280 is associated only with one gene but in *Medline* it has only 52 documents.

V. DISCUSSION

Table 2 shows the results of the 40 document setting (which includes 2 diseases, 20 abstracts each). On average, in these 10 experiments (Table 2), our technique gives ~8% higher accuracy (See Table 6). The results with Setting–2, Setting–3, and Setting–4 are included in Tables 3, 4, and 5, respectively. As can be seen in the tables, our method achieves the best results with a smaller dataset size (40 abstracts). As the number of abstracts increases, the performance improvement decreases. This indicates that our proposed technique is most effective with small numbers of documents. This is ideal for the classification of rare and relatively unknown diseases, as explained earlier. This situation could most obviously occur when a user, *e.g.* medical doctor, searches for documents on a very specific (specialized) disease (e.g. *Hematogenous metastasis ovarian cancer*). Table 6 and Figure 2 summarize the performance improvement of our method compare to the baseline. In another set of experiments, and to investigate the effectiveness of our method in a different situation, we obtained another text classification dataset: the 2007 *Medical NLP Challenge* dataset, which is typically used for general clinical text classification tasks [21]. In this dataset, documents are categorized using coding of the *ICD-9-CM* codes [17]. For 2-class classification, we separated and then labeled the dataset to distinguish non-overlapping codes. For each class pair, the first class contains all documents with a specific *ICD* code and the second class includes all the remaining documents. The results are shown in Table 7. These results confirm the effectiveness of the proposed technique with an average accuracy improvement of 3% over all four cases (this is statistically significant *$p<0.003$* with hypergeometric test). This performance improvement is similar to our disease abstract classification experiments in *Setting–4* with 500 documents (shown in Table 5). Therefore, the main contribution of this work lies in the case of a limited corpus and small document size such as is the case with disease abstracts. In such situations, the classification performance is difficult to improve with traditional classifiers, feature selection, and encoding methods. In one final evaluation batch, we extracted 200 abstracts from the disease abstract experiments shown above. Specifically, we started with the abstracts of the *Autism–Asperger* experiments.

We ran both techniques with 200 abstracts (100 for each disease), then we repeated the experiment 9 times, reducing the number of abstracts in each class (disease) by 10 (*e.g.,* the second experiment contained 180 abstracts, *...and so on.*) This was conducted to examine the behavior of our technique while varying the number of documents in a consistent setting. The results are shown in Table 8 and also illustrated in Figure 5. Our technique consistently produced an improvement in accuracy and AUC as number of documents in each classification experiment was reduced. The accuracy improvement increases from ~3% (in the case of 200 documents) to 7.6% (in the case of 40 documents). This is one more proof that the proposed technique excels in situations with a small and limited corpus. Without our technique, we experienced difficulties in (significantly) improving the classification performance across almost all learners, feature selections, and encoding methods when the corpus size is limited, e.g. with ~40 to 100 documents in both classes, which is the reason we used the *Bayesian* method after conducting 40 classification tasks of various sizes (from Table 1) and the results are shown in Table 9 and illustrated in Figure 6. Essentially, we applied our learned features in 40 classification tasks with medical abstracts of various sizes using the four machine learning methods: *Bayesian*, *SVM*, *Decision tees* (*J48*), and *Decision Forests*. These were tuned to their best performance. We counted how many times each method ranked #1 (gave best accuracy), and then how many times ranked #2, .etc. See Table 9. Bayesian showed the best overall performance in this evaluation. These results are also illustrated in Figure 6. The strength of our proposed technique comes from the way that it gives leverage to class distribution over an attribute for each attribute rather than the attribute probability distribution over each class.

TABLE 4: Performance results using Setting-3 with 200 documents including 100 disease abstracts for each disease.

Experiment (Disease-1: Disease-2)	Total Number of Attributes	Accuracy		AUC	
		Baseline	Proposed	Baseline	Proposed
Alzheimer : Parkinson	4535	0.825	0.840	0.880	0.903
Lung Cancer : Breast Cancer	4458	0.675	0.740	0.692	0.776
Lung Cancer : Breast Cancer #2	4451	0.660	0.735	0.688	0.750
Alzheimer : Parkinson #2	4528	0.825	0.849	0.880	0.910
Influenza : Measles		0.767	0.800	0.810	0.816
Cataract : Glaucoma	4440	0.720	0.772	0.764	0.834
Heart Disease : Stroke		0.750	0.763	0.782	0.798
Heart Disease : Stroke #2	4587	0.675	0.695	0.688	0.703
Anemia : Malaria	1639	0.885	0.899	0.911	0.921
Cataract : Glaucoma #2	1444	0.755	0.765	0.791	0.819
Total Number of Documents: 200 (*100 Disease-1 + 100 Disease-2*)					

TABLE 5: Performance results, Accuracy and AUC, using 500 documents in Setting 4.

Experiment (Disease-1: Disease-2)	Total Number of Attributes	Accuracy		AUC	
		Baseline	Proposed	Baseline	Proposed
Cataract : Glaucoma	7542	0.778	0.792	0.826	0.853
Anemia : Malaria	7410	0.852	0.854	0.730	0.883
Heart Disease : Stroke	7630	0.746	0.788	0.778	0.808
Influenza : Measles	6975	0.826	0.856	0.894	0.908
Alzheimer : Diabetes	7452	0.886	0.918	0.926	0.939
Autism : Asperger	2278	0.708	0.778	0.749	0.824
Dyslexia : Dyspraxia	2315	0.892	0.898	0.914	0.930
Total Number of Documents: 500 (*250 Disease-1 + 250 Disease-2*)					

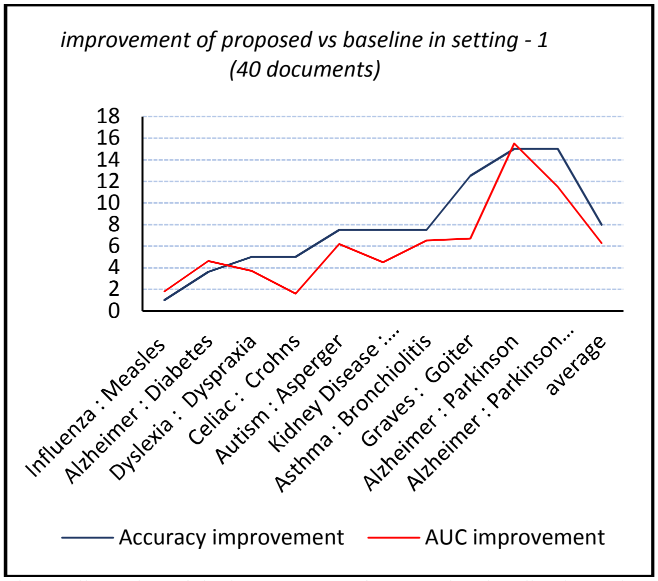

Fig. 1: Illustration of the improvement of proposed vs baseline in Setting–1 (*40 documents*).

TABLE 6: Performance improvement in classification accuracy and AUC of our technique versus the baseline.

Experimental Setting: # of doc	Improvement of proposed over baseline	
	% Accuracy	% AUC
Setting–1: 40	7.96	6.26
Setting–2: 100	4.15	4.44
Setting–3: 200	3.21	3.44
Setting–4: 500	2.80	3.39

VI. CONCLUSION

We presented a text classification technique designed for classifying *Medline* disease abstracts that can be casted as an improvement feature weighting and feature contribution to the classification task. The strength of the proposed technique is in attuning and calibrating all term features. We presented the evaluation results of the proposed technique with many text datasets and the results confirm that the technique is promising, especially with a limited corpus and relatively small number of documents. We have also shown that it outperforms the baseline technique on almost all datasets, and the baseline method, *tfidf*, is the most successful method used in text classification. Also the proposed method produced better performance in the 2007 NLP Medical Challenge benchmark dataset. With such remarkable performance results, the technique can impact and significantly contribute to the solution of other problems in bioinformatics that benefit from text classification when the available data and information is limited. This technique is ideal when we have documents and disease abstracts about very specialized disease where the number of documents is very small (less than 100). Such an effective classification system with a small dataset size can help specialized medical domain applications such as classifying medical data, cataloguing medical information, and other medical knowledge discovery applications.

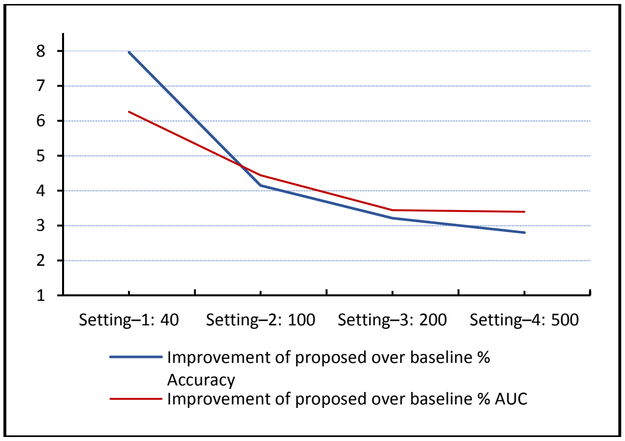

Fig. 2. Illustration of the improvement of the proposed technique in four settings compared with the baseline in accuracy and AUC.

TABLE 7: The results of our proposed technique versus the baseline in one more experiment with text documents from the *2007 NLP Medical Challenge* dataset.

Class Pair	**Number of Documents**	**Accuracy**	
		Baseline	*Proposed*
593_all	978	0.952	**0.971**
786_all	978	0.924	**0.965**
599_all	978	0.968	**0.989**
780_all	978	0.929	**0.966**
	Average	0.943	**0.973**

VII. REFERENCES

[1] V. Pekar, M Krkoska, S Staab. Feature Weighting for Co-occurrence-based Classification of Words. Proceedings of the 20th Conference on Computational Linguistics, COLING-2004

[2] George Forman, An Extensive Empirical Study of Feature Selection Metrics for Text Classification. The ACM Journal of Machine Learning Research Volume 3, 2003, pp 1289-1305

[3] Y. Saeys, I. Inza, and P. Larranaga. A review of feature selection techniques in bioinformatics. Bioinformatics, Vol. 23 no.19, 2007.

[4] H. Al-Mubaid and S.A. Umair. A New Text Categorization Technique Using Distributional Clustering and Learning Logic. IEEE Trans on Knowledge and Data Eng. vol.18, no. 9, 2006.

[5] V. Vapnik, The Nature of Statistical Learning Theory. Springer, 1995.

[6] H. Al-Mubaid, Data mining and analysis of material data using feature clustering for radiation shielding, proc. of Int'l conf on Comp. and App. in Industry and Eng. CAINE-2010, Nov. 2010.

[7] H. Al-Mubaid and S. Gungu. A Learning Based Approach for Biomedical Word Sense Disambiguation. TSWJ journal, vol. 2012, PMID: 22666174; 2012.

[8] H. Al-Mubaid. A Learning-Classification Based Approach for Word Prediction. International Arab Journal on Information Technology IAJIT, Vol.4 No.3, July 2007.

[9] LibSVM http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[10] N.A. Zaidi, J. Cerquides, M.J. Carman, and G.I. Webb. Alleviating Naive Bayes Attribute Independence Assumption by Attribute Weighting. JMLR 14 (2013) 1947-1988.

[11] J. Wu and Z. Cai. Attribute Weighting via Differential Evolution Algorithm for Attribute Weighted Naive Bayes (WNB). Journal of Computational Information Systems 7:5 (2011) 1672-1679.

[12] A. Koussounadis, O.C. Redfern and D.T. Jones. Methodology article Open Access Improving classification in protein structure databases using text mining. BMC Bioinformatics 10:129, 2009.

[13] CATH project webpage: {retrieved October 2015} http://bioinfadmin.cs.ucl.ac.uk/downloads/textCATH/

[14] Lichman, M. (2013). UCI Machine Learning Repository Irvine, CA: University of California, School of Information and Computer Science https://archive.ics.uci.edu/ml/datasets.html

[15] K. Rajeswari, S. Nakil, N. Patil et al., "Text Categorization Optimization By A Hybrid Approach Using Multiple Feature Selection And Feature Extraction Methods. Int'l Journal of Engineering Research and Applications, Vol. 4, 2014.

[16] National Library of Medicine, NIH, Medline database, PubMed (http://www.ncbi.nlm.nih.gov/pubmed)

[17] International Classification of Diseases (ICD). World Health Organization. Archived from the original on 12 February 2014. Retrieved 14 March 2014.

[18] Donaldson, Ian et al. PreBIND and Textomy – Mining the Biomedical Literature for Protein-Protein Interactions Using a Support Vector Machine. BMC Bioinformatics 4 (2003): 11. PMC. Web. 11 Nov. 2015.

[19] Dobrokhotov PB, Goutte C, Veuthey AL, Gaussier E. "A probabilistic information retrieval approach to medical annotation in SWISS-PROT". Stud Health Technol Inform. 2003.

[20] Wang J, Bo TH, Jonassen I, Myklebost O, Hovig E. Tumor classification and marker gene prediction by feature selection and fuzzy c-means clustering using microarray data. BMC Bioinformatics. 2003. [21] Computational Medicine Center. 2007. The Computational Medicine Center's 2007 Medical Natural Language Processing Challenge. http://www.people.vcu.edu/~btmcinnes/projects/icd9cm.html

[22] M. Bhasin and G.P. Raghava. 2004. ESLpred:SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. Nucleic AcidsRes.32,W414-9.

[23] PlantLoc: an accurate web server for predicting plant protein subcellular localization by substantiality motif, Nucleic Acids Research, 2013, 1–7

[24] N.Y. Yu, J.R. Wagner, M.R. Laird, G. Melli, S. Rey, R. Lo, P. Dao, S.C. Sahinalp, M. Ester, L.J. Foster, F.S.L. Brinkman PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes, Bioinformatics, (2010), 26 (13):1608-1615.

[25] M. Lan, C-L Tan, and H-B Low. Proposing a New Term Weighting Scheme for Text Categorization. AAAI 2006.

[26] Y. Yang and T. Joachims. Text categorization. Scholarpedia, 2008, 3(5):4242.

[27] F. Sebastiani. Machine learning in automated text categorization. ACM Computing Surveys, 34(1):1-47, 2002.

[28] Sasaki, Yutaka; Rea, Brian; Ananiadou, Sophia. Multi-topic Aspects in Clinical Text Classification. Bioinformatics and Biomedicine, 2007. BIBM 2007. IEEE International Conference, vol., no., pp.62-70, 2007.

[29] Garla, Vijay N, and Cynthia Brandt. Knowledge-Based Biomedical Word Sense Disambiguation: An Evaluation and Application to Clinical Document Classification. Journal of the American Medical Informatics Association : JAMIA 20.5 (2013): 882–886. PMC. Web. 1 Dec. 2015.

[30] W. Hersh, E. Voorhees. TREC genomics special issue overview. Journal Information Retrieval. Volume 12 Issue 1, Springer, 2009.

[31] Liu. F, Jenssen T.K, Nygaard V., Sack J., Hovig E., FigSearch: a figure legend indexing and classification system. Oxford Bioinformatics Journals, 2004.

[32] K. Kira and L. Rendell. A practical approach to feature selection. Proceedings of 9th int'l workshop on Machine learning, pp.249–256, San Francisco, CA, USA, 1992.

TABLE 8: The proposed technique exhibited consistently higher *accuracy* and *AUC* results with increasing improvement when reducing number of documents in the classification set. The best accuracy and AUC improvement occurs with when the size of the document corpus is only 40 documents (20 abstracts for each disease).

Number of documents (2 disease abstract sets)	**Accuracy**		**AUC**		**Avg. Accuracy improvement**	**Avg. AUC improvement**
	baseline	*proposed*	*baseline*	*proposed*		
200	0.735	0.768	0.804	0.833	3.3	2.9
180	0.761	0.794	0.825	0.858	3.3	3.3
160	0.838	0.876	0.864	0.902	3.8	3.8
140	0.800	0.841	0.840	0.878	4.1	3.8
120	0.775	0.818	0.819	0.858	4.3	3.9
100	0.760	0.806	0.787	0.829	4.6	4.2
80	0.775	0.830	0.830	0.876	5.5	4.6
60	0.783	0.844	0.834	0.881	6.1	4.7
40	0.858	0.934	0.849	0.910	7.6	6.1

TABLE 9: The accuracy results of four methods in 40 classification tasks (using medical disease abstracts selected from the 4 settings shown in Table 1) to evaluate the performance of four common learning method: Bayesian, SVM, Decision tees (J48), and Decision Forests. These were tuned to their best performance. We counted how many times each method ranked #1 (gave best accuracy), and then how many times ranked #2, and so on. The Bayesian method ranked #1 the most (18 times). It also ranked #2 the most, 13 times, and equal to SVM. Overall, Bayesian can produce the best performance in this task.

Rank	**Bayesian**	**SVM**	**J48-DT**	**D. Forest**
1	18	17	9	6
2	13	13	10	9
3	5	4	11	11
4	4	6	10	14
	40	*40*	*40*	*40*

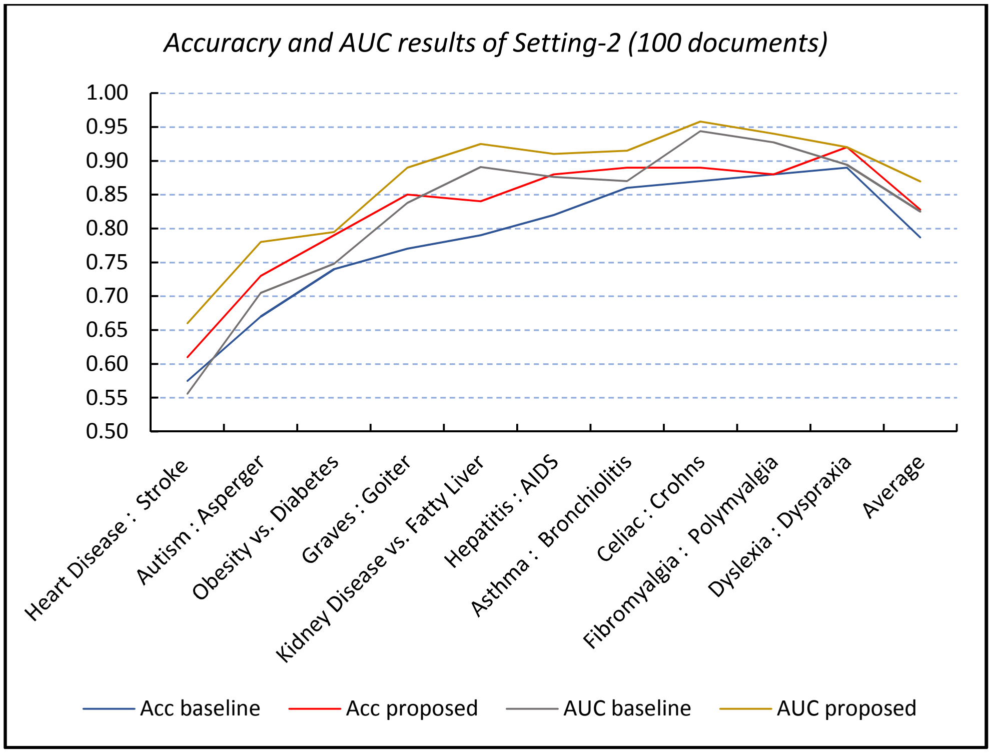

Fig. 3: Illustration of the accuracy and AUC results with Setting-2 (100 documents) taken from Table 3

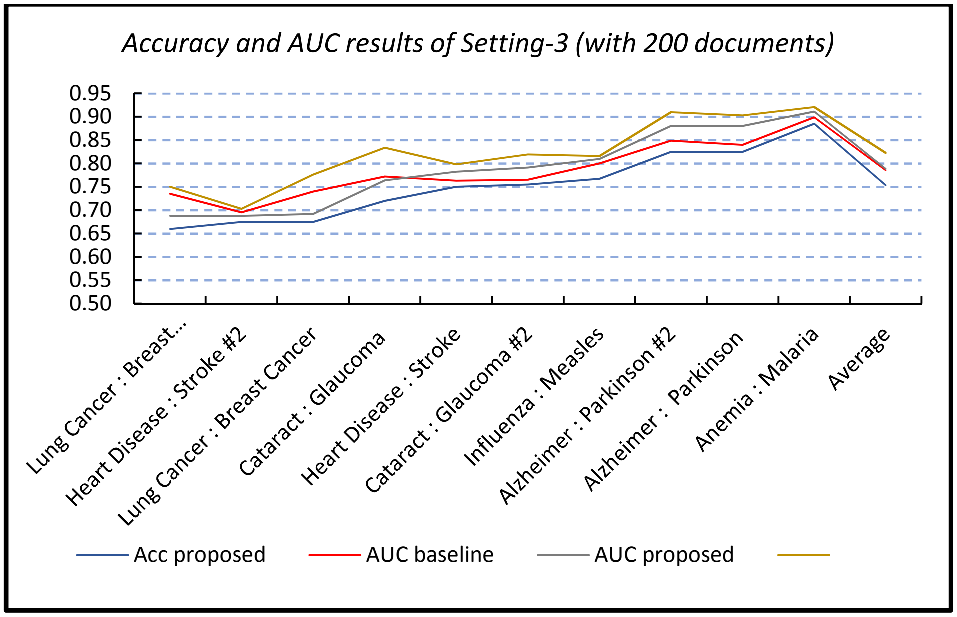

Fig. 4: Illustration of the accuracy and AUC results with Setting-3 (200 documents) taken from Table 4.

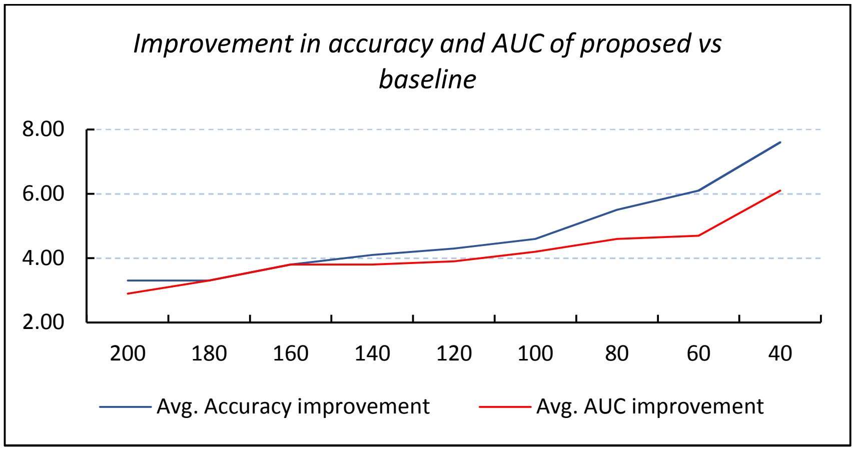

Fig. 5: Illustration of the improvement of the proposed vs baseline in terms of accuracy and AUC in the last set of experiments shown in Table 8.

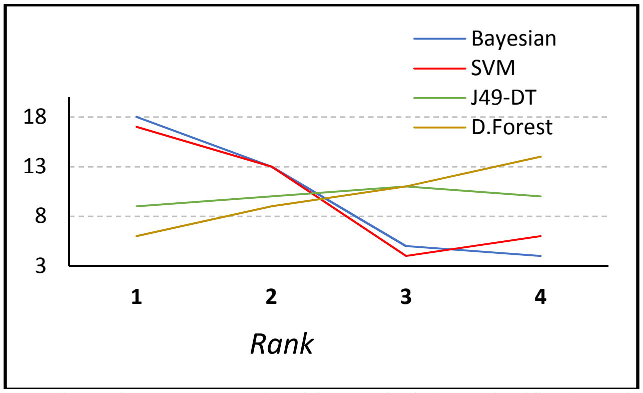

Fig. 6: This diagram shows the accuracy results of four methods in 40 classification tasks. how many times each method ranked #1, #2, #3, or #4. We notice that Bayesian ranked #1 18 times the most (also, it ranked #2 the most, 13 times, and equal to SVM). Overall, Bayesian can produce the best performance in this task.