The impact of feature selection on medical document classification

Bekir Parlak
Department of Computer Engineering
Eskişehir, Turkey
bekirparlak@anadolu.edu.tr

Alper Kursat Uysal
Department of Computer Engineering
Eskişehir, Turkey
akuysal@anadolu.edu.tr

Abstract — **Medical document classification is still one of the popular research problems inside text classification domain. Apart from some text data compiled from hospital records, most of the researchers in this field evaluate their classification methodologies on documents retrieved from MEDLINE database. OHSUMED is one of the widely used datasets containing MEDLINE documents as multi-labeled. In this study, the impact of feature selection on medical document classification is analyzed using two datasets containing MEDLINE documents. The performances of two different feature selection methods namely Gini Index and Distinguishing Feature Selector are analyzed using two pattern classifiers. These pattern classifiers are Bayesian network and C4.5 decision tree. As this study deals with single-label classification, a subset of documents inside OHSUMED and a self-constructed dataset is used for assessment of feature selection methods. Due to having low amount of documents for some categories in self-compiled dataset, only documents belonging to 10 different disease categories are used in the experiments for both datasets. Experimental results show that the combination of Distinguishing Feature Selector and Bayesian Network classifier gives more accurate results in most cases than the others.**

Keywords – text classification; medical documents; disease classification; MeSH.

I. INTRODUCTION

With the development of Internet technology, a significant increase was seen in the number of electronic documents. With this increase, automatic text classification approach has gained quite importance. The main task of automatic text classification approach is to assign the electronic documents to the appropriate classes according to their content [1]. Text classification can be used to solve a variety of problems such as the filtering of spam e-mails [2], author identification [3], classification of web pages [4], and classification of medical text documents [5][6][7].

Classification of medical abstracts is one of the main concerns inside medical text classification research field. Researches related to medical abstracts are generally carried out on MEDLINE database [8]. MEDLINE is a bibliographic database containing over 21 million documents, about 5600 medical journals. This database consists of medical abstracts in English which are assigned to some categories namely medical subject headings (MeSH). This database can be queried on internet through a search platform called PubMed [9]. Documents in MEDLINE database is indexed with corresponding relevant categories of MeSH terms by experts manually. In the literature, there exist some studies conducted on automatic classification of MEDLINE documents [6][7][10][11][12][13][14][15][16][17][18][19]. In these studies, datasets containing a certain amount of MEDLINE documents are used. The most used dataset for automatic classification of MEDLINE documents is called Ohsumed dataset. It contains medical abstracts in English for 23 types of diseases. Ohsumed, due to the structure of the MEDLINE database, is multi-label. So, it is necessary to apply multi-label classification approaches whenever a study on this dataset is performed using all documents.

In a previous study, the usage of words, medical phrases, and their combinations as features is investigated [6] for medical document classification. The results show that using combination of words and phrased as features gives slightly better classification performances than the others. In another study, multi-label classification performance based on associative classifier is examined on medical articles [10]. In another study, hidden Markov models are used for classification [14]. Besides, there exist a number of studies in the literature that ontology-based classification approaches are applied [12][16]. In a recent study, an approach using support vector machines and latent semantic indexing is applied to some datasets including the ones consisting of medical abstracts [18]. Moreover, the performances of classifiers on medical document classification is analyzed for two cases where stemming is applied and not applied [19]. Also, the impact of different text representations of biomedical texts on the performance of classification are analyzed [7].

Apart from studies that uses MEDLINE documents, there exist some medical text classification studies using data obtained from various clinics data [20][21][11][22][23][24][25]. Some of these studies concerns with medical text documents in different languages such as German [11].

In this study, the performances of two widely-known classifiers namely Bayesian networks and C4.5 decision trees are extensively analyzed using two feature selection methods on two different datasets consisting of MEDLINE documents. The first dataset is a subset of well-known OHSUMED dataset. The second one is a self-constructed dataset whose data is retrieved programmatically with querying Pubmed search platform. This dataset differs from the first one. It consists of MEDLINE documents originated from medical journals in

This work was supported by Anadolu University, Fund of Scientific Research Projects under grant number 1503F136.

Turkey. However, it has smaller amount of data than the first dataset.

Rest of the paper is organized as follows: feature extraction and selection approaches used in the study are briefly described in Section 2. Section 3 explains pattern classifiers used in this study. Section 4 presents the experimental study and results. Finally, some concluding remarks are given in Section 5.

II. FEATURE EXTRACTION AND SELECTION

A. Feature Extraction

As in most of the text classification studies, bag of words approach [1] [19] can be used for feature extraction process. In this approach, the order of terms within documents is ignored and their occurrence frequencies are used [26]. Therefore, each of the unique word in a text collection is considered as a different feature. Consequently, a document is represented by a multi-dimensional feature vector [1]. In a feature vector, each dimension corresponds to a value which is weighted by term frequency (TF), term frequency-inverse document frequency (TF-IDF) , and etc. [27].

It should also be noted that it is necessary to apply some preprocessing steps during feature extraction from text documents. Widely used preprocessing steps are "stopword removal" and "stemming". In this study, both of these two steps were applied. Porter stemming algorithm [28] was used for stemming and term frequency was used as weighting approach.

B. Feature Selection

Feature selection techniques generally fall into three categories: filters, wrappers, and embedded methods. Filter techniques are computationally fast; however, they usually do not take feature dependencies into consideration [1]. Filter-based methods are widely preferred especially for text classification domain. There is a mass amount of filter-based techniques for the selection of distinctive features in text classification. In this study, two different filter-based feature selection methods namely Gini Index (GI) and Distinguishing Feature Selector (DFS) were used. These methods are explained below in details.

1) Gini Index (GI): GI is an improved version of the method originally used to find the best split of features in decision trees [29]. It is an accurate and fast method. Its formula is as below:

$$GI(t) = \sum_{i=1}^{M} P(t|C_i)^2 \cdot P(C_i|t)^2 \qquad (1)$$

where $P(t|C_i)$ is the probability of term t given presence of class C_i, $P(C_i|t)$ is the probability of class C_i given presence of term t, respectively.

2) Distinguishing Feature Selector (DFS): DFS is one of the recent successful feature selection methods for text classification [1] whose aim is to select distinctive features while eliminating uninformative ones considering some pre-determined criteria. DFS can be expressed with the following formula:

$$DFS(t) = \sum_{i=1}^{M} \frac{P(C_i|t)}{P(\bar{t}|C_i) + P(t|\bar{C}_i) + 1} \qquad (2)$$

where M is the total number of classes, $P(C_i|t)$ is the conditional probability of class C_i given presence of term t, $P(\bar{t}|C_i)$ is the conditional probability of absence of term t given class C_i, and $P(t|\bar{C}_i)$ is the conditional probability of term t given all the classes except C_i.

III. PATTERN CLASSIFIERS

In this study, two classifiers in Weka [30] package were used programmatically. These are Bayesian Networks and C4.5 decision tree classifiers. These algorithms are explained in details below.

A. Bayesian Networks (BN)

BN is one of the methods which are used to denote modeling and state transitions [31]. BN is often used for modeling discrete and continuous variables of multinomial data. These networks encrypt the relationships between variables in the modeled data. In BN, the nodes are interconnected by arrows to indicate the direction of engagement with each other.

B. C 4.5 Decision Tree (DT)

The main purpose of the decision tree algorithms is to split the feature space into unique regions corresponding to the classes [1]. An unknown feature vector is assigned to a class via a sequence of Yes/No decisions along a path of nodes of a decision tree. C4.5 is an algorithm used to generate a decision tree and it is known as one of the successful decision tree classification algorithms.

IV. EXPERIMENTAL WORK

In this section, an in-depth investigation was carried out to measure the performance of feature selection methods and classifiers. For this purpose, combinations of feature selection methods with BN and DT classifiers were analyzed in order to determine the best combination for both of the datasets. Also, the effect of dimension reduction can be inferred according to the experimental results. In the following subsections, the utilized datasets and success measures are briefly described. Then, the experimental results are presented.

A. Datasets

In this study, two different datasets containing MEDLINE documents were used. The first one is a subset of well-known Ohsumed dataset. It consists of medical abstracts collected in 1991 related to 23 cardiovascular disease categories. As this study deals with single-label text classification, the documents belonging to multiple categories are eliminated. Also, only 10 classes are used for classification in order to make the class distribution same with the second dataset. The second dataset

is a self-constructed dataset whose data is retrieved programmatically with querying Pubmed search platform. This dataset is constructed via retrieving XML results containing medical abstracts and parsing it appropriately. The documents having multiple categories are removed from this dataset because of concerning single-label classification of medical documents. This dataset differs from the first one depending on its origins. It consists of MEDLINE documents only originated from medical journals in Turkey rather than originating from different locations. However, it has same categories with smaller amount of data than the first one. In this dataset, 10 categories having enough number of documents were used for the evaluation. The detailed information regarding those datasets is provided in Table I and Table II. In the experiments, seventy percent of documents in each class was used training and the rest was used for testing.

TABLE I. OHSUMED DATASET

Class Number	**Disesase Category**	***Number of Documents***
1	Bacterial Infections and Mycoses	631
2	Virus Diseases	249
3	Parasitic Diseases	183
4	Neoplasms	2513
5	Musculoskeletal Diseases	505
7	Stomatognathic Diseases	132
8	Respiratory Tract Diseases	634
10	Nervous System Diseases	1328
14	Cardiovascular diseases	2876
23	Pathological Conditions, Signs and Symptoms	1924

TABLE II. OHSUMED DATASET

Class Number	**Disesase Category**	***Number of Documents***
1	Bacterial Infections and Mycoses	284
2	Virus Diseases	44
3	Parasitic Diseases	116
4	Neoplasms	32
5	Musculoskeletal Diseases	140
7	Stomatognathic Diseases	39
8	Respiratory Tract Diseases	90
10	Nervous System Diseases	83
14	Cardiovascular diseases	231
23	Pathological Conditions, Signs and Symptoms	73

B. Accuracy analysis

Varying numbers of the features, which are selected by each selection method, were fed into DT and BN classifiers. In the experiments, stopword removal and stemming were applied. Widely-known Porter stemmer was carried out as stemming algorithm. In this study, GI and DFS are used as feature selection methods. Dimension reduction was carried out by constructing feature sets consisting of 300, 500, 1000, and 2000 features. Also, F-score [32] was used as success measure. This score is presented as both class specific and weighted average. Resulting F-Scores are listed in Table III and Table IV. The best ones in the results are shown as bolded.

TABLE III. RESULTS ON OHSUMED DATASET

Number of Features	**Options**				
	DFS+DT	***DFS+BN***	***GI+BN***	***GI+DT***	***Classes***
300	0,57	**0,65**	0,63	0,46	C1
	0,62	0,56	0,50	0,55	C2
	0,69	**0,77**	0,76	0,62	C3
	0,83	**0,85**	0,83	0,81	C4
	0,50	**0,58**	0,50	0,42	C5
	0,35	**0,59**	0,58	0,17	C7
	0,59	**0,62**	0,61	0,52	C8
	0,65	**0,67**	0,65	0,57	C10
	0,86	**0,86**	0,84	0,84	C14
	0,45	**0,47**	0,44	0,38	C23
Weighted Average	0,69	**0,71**	0,68	0,64	
500	0,55	**0,67**	0,66	0,51	C1
	0,58	0,52	0,53	0,50	C2
	0,69	0,74	**0,78**	0,70	C3
	0,84	**0,84**	0,82	0,80	C4
	0,46	**0,57**	**0,57**	0,44	C5
	0,24	0,56	**0,57**	0,32	C7
	0,62	**0,62**	0,60	0,48	C8
	0,66	**0,66**	0,65	0,58	C10
	0,85	**0,86**	0,84	0,82	C14
	0,44	**0,45**	**0,45**	0,41	C23
Weighted Average	0,69	**0,70**	0,69	0,64	
1000	0,55	**0,72**	0,68	0,50	C1
	0,58	0,52	0,51	0,50	C2
	0,71	**0,73**	0,70	0,68	C3
	0,83	**0,83**	0,82	0,82	C4
	0,47	**0,58**	**0,58**	0,46	C5
	0,27	**0,55**	0,51	0,24	C7
	0,61	**0,63**	0,62	0,54	C8
	0,63	**0,7**	0,68	0,58	C10
	0,84	**0,86**	0,85	0,81	C14
	0,43	**0,47**	0,46	0,41	C23
Weighted Average	0,68	**0,71**	0,70	64	
2000	0,51	**0,72**	**0,72**	0,50	C1
	0,61	0,5	0,5	0,56	C2
	0,67	**0,74**	0,73	0,65	C3
	0,82	**0,84**	0,83	0,81	C4
	0,46	0,57	**0,58**	0,46	C5
	0,14	0,51	**0,52**	0,24	C7
	0,61	0,62	**0,63**	0,53	C8
	0,63	**0,71**	0,7	0,64	C10
	0,84	**0,86**	0,85	0,83	C14
	0,42	**0,47**	**0,47**	0,40	C23
Weighted Average	0,67	**0,72**	0,71	0,65	

TABLE IV. RESULTS ON SELF-CONSTRUCTED DATASET

Number of Features	Options				
	DFS+DT	*DFS+BN*	*GI+BN*	*GI+DT*	*Classes*
300	**0,81**	**0,81**	0,79	0,8	C1
	0,67	0,42	0,44	0,57	C2
	0,86	0,72	0,72	0,84	C3
	0,63	0,31	0,31	0,57	C4
	0,62	0,76	**0,77**	0,68	C5
	0,67	**0,7**	**0,7**	0,67	C7
	0,74	0,55	0,59	0,6	C8
	0,27	0,43	**0,39**	0,3	C10
	0,72	**0,88**	0,87	0,69	C14
	0,53	0,49	0,52	**0,56**	C23
Weighted Average	**0,70**	**0,70**	0,69	0,68	
500	0,8	**0,81**	**0,81**	**0,81**	C1
	0,59	0,31	0,31	**0,62**	C2
	0,84	0,77	0,77	**0,88**	C3
	0,63	0,31	0,31	**0,63**	C4
	0,67	**0,77**	**0,77**	0,71	C5
	0,67	**0,75**	**0,75**	0,67	C7
	0,63	0,58	0,57	0,56	C8
	0,37	0,45	**0,46**	0,36	C10
	0,67	**0,89**	0,88	0,7	C14
	0,57	0,53	0,53	**0,59**	C23
Weighted Average	0,68	**0,71**	**0,71**	0,70	
1000	**0,82**	0,81	0,81	0,8	C1
	0,58	0,31	0,31	0,54	C2
	0,83	0,77	0,77	**0,85**	C3
	0,5	0,31	0,31	**0,57**	C4
	0,73	0,77	**0,77**	0,71	C5
	0,67	**0,75**	**0,75**	0,6	C7
	0,67	0,58	0,58	0,65	C8
	0,51	0,45	0,45	**0,51**	C10
	0,7	**0,89**	**0,89**	0,73	C14
	0,59	0,53	0,53	0,56	C23
Weighted Average	**0,71**	**0,71**	**0,71**	**0,71**	
2000	**0,82**	0,81	0,81	0,8	C1
	0,54	0,31	0,31	**0,56**	C2
	0,87	0,77	0,77	**0,93**	C3
	0,67	0,43	0,31	0,63	C4
	0,69	**0,78**	0,77	0,74	C5
	0,63	**0,78**	0,75	0,67	C7
	0,58	0,57	0,58	**0,59**	C8
	0,58	0,46	0,45	0,46	C10
	0,73	**0,89**	**0,89**	0,68	C14
	0,39	**0,53**	**0,53**	0,5	C23
Weighted Average	0,70	**0,71**	**0,71**	0,70	

Considering the highest weighted average F-scores, in most cases, DFS is superior to GI. In a small part of experiments, DFS and GI give similar results on both of the two datasets. It should be noted that DFS seems more successful when the feature size is low. Besides, in spite of originated from different sources and having different class-based distributions, the maximum classification performances obtained on these two datasets are similar. BN classifier is more successful than DT classifier in most of the cases.

Considering class based F-scores, classification performances obtained on neoplasms (C4) and cardiovascular diseases (C14) categories are generally higher than the others for the first dataset. This may be due to having high amount of training instances for these two categories. For self-constructed dataset, classification performances obtained on parasitic diseases (C3) and cardiovascular diseases (C14) categories are generally higher than the others. In this case, these are not the classes with maximum number of documents. This situation may be caused by having small amount of data for most of the categories. Also, for most of the class-based F-scores, combination of DFS and BN seems better than the other ones.

V. CONCLUSIONS

In this study, the performances of two widely-known classifiers are extensively analyzed using two different feature selection methods. This analysis is realized on two different datasets consisting of MEDLINE documents. In the experiments, stopword removal and stemming as preprocessing steps are applied. Experimental results show that the most successful setting is the combination of Bayesian Network classifier and Distinguishing Feature Selector. As a future work, a new dataset containing Turkish versions of the documents in the self-constructed dataset may be compiled and classification performances of these two datasets having same documents in different languages can be extensively analyzed.

VI. REFERENCES

[1] Uysal, A. K., & Gunal, S. (2012). A novel probabilistic feature selection method for text classification. Knowledge-Based Systems, 36, 226-235.

[2] Idris, I., Selamat, A., Nguyen, N. T., Omatu, S., Krejcar, O., Kuca, K., & Penhaker, M. (2015). A combined negative selection algorithm–particle swarm optimization for an email spam detection system. Engineering Applications of Artificial Intelligence, 39, 33-44.

[3] Zhang, C., Wu, X., Niu, Z., & Ding, W. (2014). Authorship identification from unstructured texts. Knowledge-Based Systems.

[4] Özel, S. A. (2011). A Web page classification system based on a genetic algorithm using tagged-terms as features. Expert Systems with Applications,38(4), 3407-3415.

[5] Garla, V., Taylor, C., & Brandt, C. (2013). Semi-supervised clinical text classification with Laplacian SVMs: An application to cancer case management. Journal of biomedical informatics, 46(5), 869-875.

[6] Yetisgen-Yildiz, M., & Pratt, W. (2005). The effect of feature representation on MEDLINE document classification. In AMIA annual symposium proceedings(Vol. 2005, p. 849). American Medical Informatics Association.

[7] Yepes, A. J. J., Plaza, L., Carrillo-de-Albornoz, J., Mork, J. G., & Aronson, A. R. (2015). Feature engineering for MEDLINE citation categorization with MeSH.*BMC bioinformatics*, *16*(1), 1.

[8] MEDLINE [http://www.nlm.nih.gov/databases/databases_medline.html] . Accessed 2015.

[9] Pubmed [http://www.ncbi.nlm.nih.gov/pubmed]. Accessed 2015.

[10] Rak, R., Kurgan, L. A., & Reformat, M. (2007). Multilabel associative classification categorization of MEDLINE articles into MeSH keywords. IEEE engineering in medicine and biology magazine, 26(2), 47.

[11] Spat, S., Cadonna, B., Rakovac, I., Gutl, C., Leitner, H., Stark, G., & Beck, P. (2007). Multi-label text classification of German language medical documents. In Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems (p. 2343). IOS Press.

[12] Camous, F., Blott, S., & Smeaton, A. F. (2007). Ontology-based MEDLINE document classification. In Bioinformatics Research and Development (pp. 439-452). Springer Berlin Heidelberg.

[13] Poulter, G. L., Rubin, D. L., Altman, R. B., & Seoighe, C. (2008). MScanner: a classifier for retrieving Medline citations. BMC bioinformatics, 9(1), 108.

[14] Yi, K., & Beheshti, J. (2008). A hidden Markov model-based text classification of medical documents. Journal of Information Science.
[15] Frunza, O., Inkpen, D., Matwin, S., Klement, W., & O'blenis, P. (2011). Exploiting the systematic review protocol for classification of medical abstracts. Artificial intelligence in medicine, 51(1), 17-25.
[16] Dollah, R. B., & Aono, M. (2011). Ontology based Approach for Classifying Biomedical Text Abstracts. International Journal of Data Engineering (IJDE),2(1), 1-15.
[17] Albitar, S., Espinasse, B., & Fournier, S. (2014, March). Semantic Enrichments in Text Supervised Classification: Application to Medical Domain. In The Twenty-Seventh International Flairs Conference.
[18] Uysal, A. K., & Gunal, S. (2014). Text classification using genetic algorithm oriented latent semantic features. Expert Systems with Applications, 41(13), 5938-5947.
[19] Parlak, B., & Uysal, A. K. (2015, May). Classification of medical documents according to diseases. In *Signal Processing and Communications Applications Conference (SIU), 2015 23th* (pp. 1635-1638). IEEE.
[20] Pakhomov, S. V., Buntrock, J. D., & Chute, C. G. (2006). Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. Journal of the American Medical Informatics Association, 13(5), 516-525.
[21] Van Der Zwaan, J., Sang, E. T. K., & de Rijke, M. (2007). An experiment in automatic classification of pathological reports. In Artificial Intelligence in Medicine (pp. 207-216). Springer Berlin Heidelberg.
[22] Waraporn, P., Meesad, P., & Clayton, G. (2010). Ontology-supported processing of clinical text using medical knowledge integration for multi-label classification of diagnosis coding. arXiv preprint arXiv:1004.1230.
[23] Boytcheva, S. (2011, September). Automatic Matching of ICD-10 codes to Diagnoses in Discharge Letters. In Proceedings of the Workshop on Biomedical Natural Language Processing, Hissar, Bulgaria (pp. 11-18).
[24] CEYLAN, N. M., ALPKOÇAK, A., & ESATOĞLU, A. E (2012). Tıbbi Kayıtlara ICD-10 Hastalık Kodlarının Atanmasına Yardımcı Akıllı Bir Sistem.
[25] Arifoğlu, D., Deniz, O., Aleçakır, K., & Yöndem, M. (2014). CodeMagic: Semi-Automatic Assignment of ICD-10-AM Codes to Patient Records. In Information Sciences and Systems 2014 (pp. 259-268). Springer International Publishing.
[26] Uysal, A. K., Gunal, S., Ergin, S., & Gunal, E. S. (2012, April). Detection of SMS spam messages on mobile phones. In Signal Processing and Communications Applications Conference (SIU), 2012 20th (pp. 1-4). IEEE.
[27] C.D. Manning, P. Raghavan, H. Schutze. Introduction to information retrieval Cambridge University Press, New York, USA (2008)
[28] M.F. Porter, "An Algorithm for Suffix Stripping," Program, vol. 14, p. 130-137, 1980.
[29] W. Shang, H. Huang, H. Zhu, Y. Lin, Y. Qu, Z. Wang, A novel feature selection algorithm for text categorization, Expert Systems with Applications 33 (1) (2007) 1–5.
[30] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.
[31] Ian H Witten and Eibe Frank, Data Mining: Practical Machine Learning Tools and Techniques, Jim Gray, Ed. San Fransisco: Morgan Kaufmann Publishers, 2005.
[32] Goutte C, Gaussier E (2005) A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In: Proc. Europe Conf. Information Retrieval Research, pp. 345–359