International Journal of
Intelligent Systems and
Applications in Engineering
ISSN:2147-6799 www.atscience.org/IJISAE

Advanced Technology and Science

Original Research Paper

An Efficient Document Categorization Approach for Turkish Based Texts

Sevinç İlhan Omurca*[1], Semih Baş[2], Ekin Ekinci[1]

Received 14th September 2014, Accepted 18th December 2014

Abstract: Since, it is infeasible to classify all the documents with human effort due to the rapid and uncontrollable growth in textual data, automatic methods have been approached in order to organize the data. Therefore a support vector machine (SVM) classifier is used for text categorization in this study. In text categorization applications, the text representation process could take a huge computation time on weighting the huge size of terms. So far, lexicons that contain less number of terms are used for the solution in the literature. However it has been observed that these kinds of solutions reduce the accuracy of the text classification. In this paper, the term-document matrix is constructed as user dependent according to the purpose of classification. Since the number of terms is still relatively large, we used a hash table for efficient search of terms. Hereby an efficient and rapid TF-IDF method is introduced to construct a weight-matrix to represent the term-document relations and a study concerning classification of the documents in Turkish based news and Turkish columnists is conducted. With the proposed study, the computational time that is required for term-weighting process is reduced substantially; also 99% accuracy is achieved in determination of the news categories and 98% accuracy is achieved in detection of the columnists.

Keywords: *Document categorization, SVM, TF-IDF, User dependent term selecting, Hash table.*

1. Introduction

Due to the rapid and uncontrollable growth in textual data, especially with the domain World Wide Web (www), it is infeasible to manually classify the huge size of documents with high-dimensional text features, so the automatic methods for organizing the data are needed. Text classification is the task of assigning the documents to a set of predefined classes based on their contents. Classification of web pages, filtering of spam e-mails, categorization of topics, retrieving user reviews, author recognition are some popular application areas of text classification.

There are certainly a broad range of machine learning methods available for text classification problems in the literature. The most popular ones include regression models, probabilistic Bayesian models, decision trees, decision rule learners, K-nearest neighbors (KNN), computing with words, association rule mining and SVM. Among these methods, SVM achieves superior results in text classification and pattern recognition problems [1]. (Fabrizio Sebastiani, 2005) also emphasized SVM classifier in his review paper of text categorization because of its best performance in comparative text categorization experiments so far.

Here some of the approaches and techniques have been applied recently in the field of text classification are referred. (Zhang et al; 2008) investigate the effectiveness of using multi-words for text classification with SVM and also effectiveness of linear kernel and polynomial kernel in SVM comparatively. (Li et al; 2011) proposed a hybrid algorithm that combines SVM and KNN and overcomes the drawbacks of sensitive to noises of SVM and low efficiency of KNN. (Sun et al; 2009) realized a comparative study on the effectiveness of strategies addressing imbalanced text classification using SVM and make a survey on the techniques proposed for imbalanced classification. (Miao et al; 2009) proposed a hybrid algorithm which is based on variable precision rough set and KNN to overcome their weaknesses. (Shi et al; 2011) studied semi-supervised text classification; they tried to learn from positive data without negative data and also with the help of unlabeled data. They use SVM, Naive Bayes and Rocchio as classifiers to construct a set of classifier. (Mitra et al; 2007) proposed a least square support vector machine (LS-SVM) that classifies noisy document titles and the proposed system was compared with KNN and Naive Bayes. It was observed that LS-SVM with LSI based classifying agents improves text classifying performance significantly. (Lo, 2008) proposed an auto mechanism to classify customer messages based on the techniques of text mining such as dictionary approach or TF-IDF and SVM then exceeded 83-89% success in classifying. (Rajan et al; 2009) proposed an ANN model for the classification of Tamil language documents and the model achieved 93.33% accuracy. (Zhang et al; 2013) used Rough Set which is based on Rough Set decision making approach for classifying texts which are not easily classified with classical methods. They used CEI for performance evaluation. (Adeva et al; 2014) studied with SVM, Naïve Bayes, KNN and Rocchio for medical-domain texts. They combined these algorithms with 7 different feature selection algorithms and different number of features and used 3 different document sections. (Lee et al; 2012) proposed a new approach, called as Euclidean-SVM. In training phase they used SVM and in classification phase they used Euclidean distance function instead of optimal hyper-plane.

Due to the literature, there are only a few text classification approaches that have been applied in Turkish documents.

[1] Kocaeli University, Faculty of Engineering, Computer Engineering Department Umuttepe Campus, Kocaeli – 41380, Turkey
[2] Tubitak Marmara Research Center Technology Free Zone, IBTECH, Kocaeli – 41470, Turkey
** Corresponding Author: Email: silhan@kocaeli.edu.tr*

Note: This paper has been presented at the International Conference on Advanced Technology&Sciences (ICAT'14) held in Antalya (Turkey), August 12-15, 2014.

(Kılıçaslan et al; 2009) explored machine learning models such as Navie Bayes, KNN, decision trees, SVM and voted perceptron for pronoun resolution in Turkish. (Çıtlık and Güngör, 2008) employed SVM and boosting classifiers in spam filtering and achieved high accuracies. (Özyurt and Köse, 2010) studied chat mining. They used Naive Bayes, KNN and SVM to classify Turkish chat conversation texts and achieved 90% accuracy in determination of subject. (Özgür et al; 2004) proposed an anti spam filtering based on Artificial Neural Networks and Bayesian Networks. They tested the system with 750 e-mails and achieved 90% accuracy. (Alparslan et al; 2011) proposed a hybrid system for document classification that considers SVM and adaptive neuro-fuzzy classifier and 96.67% accuracy was achieved. (Uysal and Gunal, 2014) proposed to show impact of preprocessing on text classification. To this end, they used SVM to classify Turkish and English news and e-mails.

In this paper, we have applied a supervised machine learning method in order to classify the Turkish news and also predict the columnists of newspaper articles. There are not many work have been done in Turkish news classification or author detection. (Türkoğlu et al; 2007) identified the author of an unauthorized document by using n-grams and determined the most success classifiers were SVM and Multi Layer Perceptron (MLP). An average accuracy of 88.9% was achieved by SVM. In the current method by using the TF-IDF term weighting method and SVM classifier, a success rate of 96.4% and a lower time complexity are obtained. Thus, it is concluded that great time savings are possible without decreasing the accuracy level.

Our study has two main phases, the first one is text representation phase that is realized by TF-IDF method and the second one is the text classification phase that is realized by a SVM classifier. In text representation process, the huge size of terms entire dataset namely huge amount of feature set causes huge computation time on weighting these terms [21]. In the text representation phase of our application, unlike from the other applications in the literature, the words that are inefficient for classification are subtracted to reduce the term space. The subtracting process is realized by the user due to the characteristics and purposes of classification task. Namely, if sentiment classification of the textual data will be realized then the verbs would be so important even the proper names would not. On the other hand, if the category of documents will be estimated then the proper names would be so important. Or generally, the conjunctions have less importance in text mining independently from the classification task. These kinds of determinations about the term selecting process must be done with the expert persons on the classification tasks.

The rest of the paper is organized as follows. Section 2 gives an overview of the TF-IDF and SVM methods. Section 3 discusses the experimental setup. Section 4 shows the results of experiments and section 5 gives the concluding remarks.

2. Brief Overview

2.1. TF-IDF

In text classification problems, for most of the training algorithms, a document should be represented as a vector of numbers. A method called term frequency (TF) and inverse document frequency (IDF) are used to represent text with vector space model. There is an extension of term frequency inverse document frequency (TF-IDF) developed from IDF which is proposed by ([22], [23]) and expresses that a term which appears in many documents is not a good discriminator and should be given less weight than another term which appears in few documents [24]. Intuitively, this method determines how relevant a given word is in a particular document. Terms that are common in a small group of document-set tend to have higher TF-IDF numbers than common terms such as prepositions or articles.

Assume there are N documents in the collection, t_i denotes term i and occurs n_i of documents. Then inverse document frequency is formulized as in (Equation.1).

$$IDF(t_i) = \log \frac{N}{n_i} \tag{1}$$

In text classification models, a text can be defined as a term matrix, $D = \lfloor d_{ij} \rfloor_{m \times n}$, where n denotes the number of documents, m denotes the number of different terms and d_{ij} denotes the weight value of the term t_i in document d_j. TF-IDF method expressed by (Equation.2) is used to compute the term weight values. Where tf_{ij} indicates the frequency of the term i in the document j [25].

$$dij = TF_{ij} \times IDF_i = \frac{TF_{ij} \times \log_2 (N/n_i + 0.01)}{\sqrt{\sum_{j=1}^{m} (TF_{ij} \times \log_2 (N/n_i + 0.01))^2}} \tag{2}$$

TF-IDF formulation is used to measure the discrimination or importance value of a term in the document collection [26]. However there is an important criticism of using this method for text representation. This comes from the huge dimensionality of term-document matrix, resulting in that it causes a huge computation time on weighting these terms [21]. High dimensionality of the feature space is also addressed as the major difficulty of text categorization problems. The classification accuracy is directly connected with how much of the document terms can be reduced without losing useful information in category representation [27]. Consequently a powerful test representation implementation would not only decrease the computational time for the task but also improve the accuracy of the classification task.

2.2. SVM

SVM is a computational learning method uses machine learning theory presented and developed by [28]. (Joachims, 1998) was firstly proposed SVM for text classification tasks and just it is clearly known that, SVM is one of the most important learning algorithms for text classification due to its robustness on high dimensional spaces [30].

In SVM, original input space is mapped into high dimensional feature space and in this space there are many hyper-planes (linear classifiers) that separate the data. The optimal hyper-plane among them that achieves maximum separation is determined by optimization theory to maximize the generalization ability of the classifier [31].

A training data set represented by n-dimensional input $x_i \in R^n$, $i = 1,\dots,l$ and l is the number of samples that belong to target classes $y_i \in \{1,-1\}$. A hyper plane $f(x) = 0$ that separates the data is tried to find.

$$f(x) = w \cdot x + b = \sum_i^n w_i \cdot x_i + b = 0 \tag{3}$$

where $w = (w_1,\dots,w_n)$ and $b \in R$. The aim is correctly classifying the data and a distinctly separating hyper-plane satisfies these conditions.

$$y_i(x_i \cdot w + b) \geq 1, i = 1, \ldots, l \quad (4)$$

$$f(x) = w \cdot x_i + b \geq 1, y_i = +1 \quad (5)$$

$$f(x) = w \cdot x_i + b \leq 1, y_i = -1 \quad (6)$$

Among all possible hyper-planes SVM selects the optimal separating hyper-plane that creates the maximum margin. The optimal hyper plane can be found by solving a quadratic optimization problem in (Equation 7). ξi is slack variable represents noise and C is error penalty determines the trade-off between model complexity and loss function.

Minimize: $$\phi(w\xi) = \frac{1}{2(w \cdot w)} + C(\Sigma_{i=1}^{l}\xi_i) \quad (7)$$

Subject to: $$y_i(x_i \cdot w + b) \geq 1 - \xi_i, i = 1, \ldots, l \quad (8)$$

For simplification of the calculations, the optimization problem has been converted to Lagrange dual problem with Kuhn-Tucker conditions.

$$\Sigma_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_i\sum_j \alpha_i \alpha_j y_i y_j K(x_i, x_j) \quad (9)$$

Subject to: $$\sum_{i=1}^{l}\alpha_i y_i = 0, 0 \leq \alpha_i \leq C, i = 1, \ldots, l \quad (10)$$

$K(x_i, x_j)$ is the inner product $\left\langle \phi(x_i)\phi(x_j) \right\rangle$ in feature space and called as kernel function.

$$K(x_i, x_j) = \left\langle \phi(x_i)\phi(x_j) \right\rangle \quad (11)$$

ϕ is a mapping from X to inner product feature space F. In practice ϕ and F are derived from the definition of kernel function. There are different kernel functions for SVM. (Joachims, 2002) and (Dumais et al; 1998) reported an important finding in text classification that linear SVM performs better than nonlinear SVM so in this paper a linear kernel function is used for SVM. The other common kernel functions are as follows and called Polynomial Kernel, Radial Basis Kernel and Sigmoid Kernel Function respectively.

$$K(x_i, x_j) = \left[\left(x_i \cdot x_j + 1\right)\right]^q \quad (12)$$

$$K(x_i, x_j) = \exp\left(-\frac{\left\|x_i - x_j\right\|}{2\alpha^2}\right) \quad (13)$$

$$K(x_i, x_j) = \tanh\left(v\left(x_i, x_j\right) + C\right) \quad (14)$$

The final mapping function $f(x)$ between the input variable space and the desired output variable can be expressed in terms of the SVs (training examples) as follows:

$$f(x) = \Sigma_{i,j=0}\alpha_i y_i K(x_i, x_j) + 1 \quad (15)$$

where x_i, x_j are SVs for class 1 and class 2, respectively [31].

3. Experimental Setup

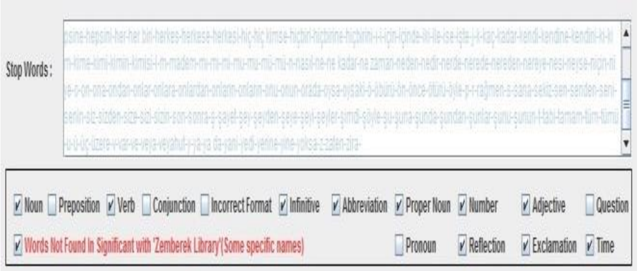
3.1. Text Collection and Pre-processing

The first phase of the text classification model is the pre-processing phase which includes elimination of stop words, determination of the terms and stemming. Stop words elimination filters out the words that are not relevant in the analysis of documents and usually consist of articles, pronouns, prepositions, interjections among others.

For the pre-processing step of the application, initially, we eliminate the stop words from the document set directly. The stop word list including about 223 words is obtained from [33]. Thus the computation complexity for multiword representation can be reduced by stop word elimination because they usually have high frequency in documents. Then, the user selected terms are also eliminated from the document set. For the classification experiments in this study, different kinds of terms like noun, preposition, verb, conjunction, incorrect format, infinitive, abbreviation, proper noun, number, adjective, question, pronoun, reflection, exclamation, time and words not found significant with Zemberek Library (some specific shortenings) can be chosen for the elimination. After all, the document list that composed of selected terms is handled to find out the root forms of words by a comprehensive Turkish stemming library Zemberek [34] in order to reduce the number of terms needed to represent the document collection. After the pre-processing step, term-document matrix contains in its cells the importance of terms in the document set have been constructed.

Differently from the current studies on the subject, in this study a user dependent term selection and weighting method is used. The proposed study allows the user to eliminate the words by the term selection among different function words like prepositions, pronouns, conjunctions and also among different content words like nouns, adjectives, verbs, shortenings or proper names. Every different task of text categorization may require different kinds of function and content term analysis. For instance, while classifying the documents that include mathematical information the numbers are so important; however the numbers do not have any importance for sentiment classification. In category determination, the proper names may be so important for defining the magazine category; however in sentiment analysis they are not so important, the adjectives are more important. In author detection, frequently used articles, prepositions may be helpful. In brief the right terms should be chosen for the right task requires a perception of the nature of text categorization task [2]. In this regard, the proposed term selection part is an effective factor for achieving high classification accuracy rates. The main contribution of the study is, with developed software tool the users can chose which kinds of terms must be evaluated and which kind of terms are redundant for text classification process. In other words the chosen terms are not weighted and also evaluated. Thus the term-document matrix space is intelligently decreased with respect to text categorization task.

The stop words and the terms that can be chosen by the users for constructing term-document matrix are shown in (Figure.1).



Stop Words : psine-hepsini-her-her biri-herkes-herkese-herkesi-hiç-hiç kimse-hiçbiri-hiçbirine-hiçbirini-ı-i-için-içinde-iki-ile-ise-işte-j-k-kaç-kadar-kendi-kendine-kendini-ki-ki
m-kime-kimi-kimin-kimisi-l-m-madem-mi-mı-mi-mu-mu-mü-mü-n-nasıl-ne-ne kadar-ne zaman-neden-nedir-nerde-nerede-nereden-nereye-nesi-neyse-niçin-ni
ye-o-on-ona-ondan-onlar-onlara-onlardan-onların-onların-onu-onun-orada-oysa-oysaki-ö-öbürü-ön-önce-ötürü-öyle-p-r-rağmen-s-sana-sekiz-sen-senden-seni-
senin-siz-sizden-size-sizi-sizin-son-sonra-ş-şayet-şey-şeyden-şeye-şeyi-şeyler-şimdi-şöyle-şu-şuna-şunda-şundan-şunlar-şunu-şunun-t-tabi-tamam-tüm-tümü
-u-ü-üç-üzere-v-var-ve-veya-veyahut-y-ya-ya da-yani-yedi-yerine-yine-yoksa-z-zaten-zira-

☑ Noun □ Preposition ☑ Verb □ Conjunction □ Incorrect Format ☑ Infinitive ☑ Abbreviation ☑ Proper Noun ☑ Number ☑ Adjective □ Question
☑ Words Not Found In Significant with 'Zemberek Library'(Some specific names) □ Pronoun ☑ Reflection ☑ Exclamation ☑ Time

Figure 1. Stop words and redundant words for the text analysis application task.

3.2. Term Weighting

In text mining, the term-document matrix is mostly weighted by TF-IDF method. In conventional TF-IDF method, how many times each term appears in document set is calculated. The major difficulty here is the high computational time caused by high dimensionality of the terms. In text mining, supervised linear feature extraction methods may be used to reduce the feature dimensionality [35]. When the relevant literature is analyzed it is seen that, the high computational time problem is usually solved by linear discriminant analysis (LDA) or any of the supervised linear feature extraction methods, in this study, without any need to reducing the term-document matrix dimension, the term frequency determination process is accelerated. This is achieved by combining the proposed user dependent term selection method with hashing method.

Since the number of unique terms in document set is relatively large, a hash table is built and used for efficient searching. The hash table consists of <key, value> pairs that are the unique terms appear in the document set and their appearing frequencies respectively. The keys represent the domain dependent and user selected unique terms; the values represent the number of documents that contain these keys. The TF-IDF values are easily computed by configuring the hash table term-based. When a term is appeared in the first document, it has been added to hash table as a key and also the frequency of it as a value. After that, while the term frequencies are been calculated for the second document, if the same term appears in this document again, that means, this term was already added to hash table and it can be easily reached by the key value. Thus, instead of searching the term in a list structure, it has been reached directly by the generated hash code. The list structure typically indexed with integer numbers, while hash table indexed with a word.

Hash structure can be very efficient for processing large scaled data, because the time to locate a value on a hash table is absolutely independent of its size. The length of the frequency list for each term is the index of the document this term is last occurred in. By this means, it saves us to make unnecessary computation loops on document set such as for a term which is only occurred in the first document. As an example the first document content is like "… the clustering application is …" and the second document is like "... the next word in the next application …" the number of documents is denoted n. Then our hash table structure for this example is as in (Table.1).

Table 1. Hash table structure

Key	Document 1	Document 2	...	Document n
	(Value list)	(Value list)		(Value list)
clustering	1	0	...	...
application	1	1	...	...
next	0	2	...	...
word	0	1	...	...

4. Classification Results

4.1. Text Collection and Pre-processing

In this study, two Turkish text datasets that taken from a natural language research group of Yıldız Technical University [36] in Turkey are used in order to examine the performance of the proposed document categorization system. The first sample data set contains 10 different columnists for each of 9 different authors. The second data set contains 150 different documents for each of 5 different news groups that have different subject in each such as economy, magazine, medical, politics and sports. In machine learning techniques, the ratio between the training data and the test data is recommended as 75% and 25% of all data respectively [37]. Accordingly, for the first dataset, 63 documents have been used for training and 27 texts for testing. The second dataset is split into 560 texts for training and 190 texts for testing. Thus different parts of the whole data are treated as training and test examples for SVM learning. Once the training phase is completed, the SVM model can be able to classify some unknown text data.

To evaluate the proposed document categorization system, four kinds of classical evaluation measures constantly used in document categorization, precision, recall, F-measure and accuracy are adopted for the experiments. Precision is a measure of the ability of a classification model to present only relevant items. Recall is a measure of the ability of a classification model to present all relevant items. F-measure is the weighted harmonic mean of precision and recall.

4.2. Experimental Results and Analysis

In this study, Java programming language is used to develop a document categorization application. We run the experiments on an Intel Core 2 duo (2.27 GHz) PC with 2GB Ram. First, we examine the proposed document categorization model on author categorization dataset. For author categorization process, the total meaningful words in training text are ranked as 3743 and the training time of the classifier is measured approximately 0.12 seconds. After the training phase, the test examples are uploaded to software and they classified by the SVM model.

For each author in this dataset, the number of training, test and misclassified examples are also shown in (Table.2).

Table 2. Classification performance due to the selected terms

Author	Training	Test	Misclassified
Doğan Hızlan	7	3	0
Erkan Çelebi	7	3	0
Ercan Mumcu	7	3	0
Ertuğrul Özkök	7	3	1
Ertuğrul Sağlam	7	3	0
Fatih Altaylı	7	3	0
Gündüz Tezmen	7	3	0
Pakize Suda	7	3	1
Serdar Turgut	7	3	0

In text mining, the importance of the terms chances due to document categorization task [2]. To demonstrate this point, the SVM classifier runs for two different user selected term sets for weighting. The first set consists of noun, abbreviation, proper noun, adjective, reflection, exclamation and the second set consists of noun, verb, infinitive, abbreviation, proper noun, adjective, reflection, exclamation and time. According to these two sets the classification results are shown in (Table.3).

Table 3. Classification performance due to the selected terms

Selected Terms	Precision %	Recall%	F-measure %	Accuracy%
1st set	91	91	91	98
2nd set	89	92	90	97

In the first case the SVM classifier achieves 98% accuracy by weighting 3743 words, but in the second case, the SVM classifier achieves 97% accuracy by weighting 4358 words. Despite, the number of meaningful terms used for constructing SVM classifier is decreased, the accuracy of classifier is increased in consequence of selecting the right dimensions for the right task. Briefly, the proposed text representation model increases the accuracy of SVM classifier; in addition that it decreases dimension of the term-document matrix and consequently the required classification time.

Secondly, the number of training, test and misclassified examples for each news document in the news-group dataset are shown in (Table.4). Three of the documents are misclassified among 190 test examples. Thus the classification performance is calculated as follows; the precision of the SVM classifier for this dataset is 98.6%, recall is 98%, F-measure is 98% and the accuracy is 99%.

Table 4. Classification results for newsgroup dataset

Category	Training	Test	Misclassified
Ekonomi	112	38	1
Magazin	112	38	2
Sağlık	112	38	0
Siyaset	112	38	1
Spor	112	38	0

There is another critical point of the results of this study. When the test results of the SVM classifier were evaluated, it was observed that the classification error rates are considerable small; what is more, when the misclassified documents were evaluated, it was observed that these documents do not contain sufficient distinguishing words to represent their categories.

Considering the newsgroup dataset will be descriptive for understanding the reasons of misclassifications. This data set contains several documents in five different news groups, such as economy, magazine, medical, politics and sports. When the results in Table 5 are examined, it is seen that one document in economy category and two documents in magazine category are misclassified.

First we evaluate the fifth misclassified document in economy category. It is classified as in medical category (3) by SVM classifier even though it is in economy category (1). Title of this news document is “The ministry of health’s objection towards the cord blood trade” and the document expresses the legal restrictions which have been applied due to limit the cord blood trade. It contains a lot of medical terms and the SVM classifier labeled it as in medical category. At first blush the document seemed misclassified; however, indeed the decision of the classifier is not so wrong.

Other misclassified documents are in magazine category. One of them is labeled as in sports category by SVM classifier even though it is in magazine category. When the content of the news document was evaluated, it was observed that it is about a Turkish basketball player who had played in AEK and Panatinaikos basketball teams. The document refers his attendance to the Olympiads in Athens and his fifteen days basketball camp for the national match. As a result of these reviews, the decision of the classifier is not so wrong. The other misclassified document in magazine category is about the hairs and it is labeled as in medical category.

5. Conclusion

There are two major factors that make the text classification process difficult. The first one is the problem of defining the document feature vector that better distinguish the category to which each document belongs. The second one is the problem of deciding the best learning model as document classifier. In this paper, a new approach for the first issue in Turkish based texts is directly addressed. Then a SVM classifier with a linear kernel function is implemented in order to observe the accuracy of the classification.

In text classification applications, the text representation process causes huge computation time on weighting the huge size of terms. Lexicons that contain less number of terms are used as a usual solution for this problem. In this study, distinctively from the literature, a user dependent term-document matrix is determined for text representation. The terms like noun, adjective, infinitive, verb, abbreviation, proper noun, number, reflection, exclamation, time and words not found significant with Zemberek Library are considered as the terms that are going to be evaluated. Due to the characteristics of the classification model, user can chose any of these terms to construct the representatives of the documents. While the characteristics and the purpose of the classification model changes, the important kind of terms will also be changed. For example, on the one hand the nouns may be more important for news classification model, on the other hand the verbs and adjectives may be more important for sentiment classification. In brief, user can determine the terms that will be used in term-document matrix construction, according to the purpose and characteristics of classification model.

Experiments conducted over author and newsgroup datasets and as a result of these 98% and 99% accuracy are achieved respectively. Increasing number of meaningful terms, which are used for constructing SVM, has caused a decrease in accuracy in determination of columnist. With 4358 meaningful term SVM classifier achieves 97% accuracy for author dataset. Our feature extraction strategy based on hash table consistently improves text classification in the terms of classification accuracy and computational time. As in other studies it has been clearly shown by this study, SVM achieves superior classification accuracy in classification problems.

The future work should be done on the issues of trying different document classification problems and determining the relation between the classification purpose and corresponding important terms. A self learning system should be developed for determining which kind of terms is useful for which kind of classification task.

References

[1] M. A. Kumar, and M. Gopal, "A comparison study on multiple binary-class SVM methods for unilabel text categorization," Pattern Recognition Letters, vol. 31, pp. 1437-1444, Aug. 2010.

[2] F. Sebastiani, Text Categorization, A. Zanasi, Ed. Southampton, UK: WIT Press, 2005.

[3] W. Zhang, T. Yoshida, and X. Tang, "Text classification based on multi-word with support vector machine," Knowledge-Based Systems, vol. 21, pp. 879-886, Dec. 2008.

[4] W. Li, D. Miao, and W. Wang, "Two-level hierarchical combination method for text classification," Expert Systems with Applications, vol. 38, pp. 2030-2039, Mar. 2011.

[5] A. Sun, E. Lim, and Y. Liu, "On strategies for imbalanced text classification using SVM: A comparative study," Decision Support Systems, vol. 48, pp. 191-201, Dec. 2009.

[6] D. Miao, Q. Duan, H. Zhang, and N. Jiao, "Rough set based hybrid algorithm for text classification," Expert Systems with Applications, vol. 36, pp. 9168-9174, July 2009.

[7] L L. Shi, X. Ma, L. Xi, Q. Duan, and J. Zhao, "Rough set and ensemble learning based semi-supervised algorithm for text classification," Expert Systems with Applications, vol. 38, pp. 6300-6306, May 2011.

[8] V. Mitra, C. Wang, and S. Banerjee, "Text classification: A least square support vector machine approach," Applied Soft Computing, vol. 7, pp. 908-914, June 2007.

[9] S. Lo, "Web service quality control based on text mining using support vector machine," Expert Systems with Applications, vol. 34, pp. 603-610, Jan. 2008.

[10] K. Rajan, V. Ramalingam, M. Ganesan, S. Palanivel, and B. Palaniappan, "Automatic classification of Tamil documents using vector space model and artificial neural network," Expert Systems with Applications, vol. 36, pp. 10914-10918, Oct. 2009.

[11] L. Zhang, Y. Li, C. Sun, and W. Nadee, "Rough Set Based Approach to Text Classification," in IEEE/WI/ACM International Conferences on Web Intelligence (WI) and Intelligent Agent Technology (IAT), 2013, p. 245.

[12] J. J. G. Adeva, J. M. P. Atxa, M. U. Carrillo, and E. A. Zengotitabengoa, "Automatic text classification to support systematic reviews in medicine," Expert Systems with Applications, vol. 41, pp. 1498-1508, Mar. 2014.

[13] L. H. Lee, C. H. Wan, R. Rajkumar, and D. Isa, "An enhanced Support Vector Machine classification framework by using Euclidean distance function for text document categorization," Applied Intelligence, vol. 37, pp. 80-99, July 2012.

[14] Y. Kılıçaslan, E. S. Güner, and S. Yıldırım, "Learning-based pronoun resolution for Turkish with a comparative evaluation," Computer Speech and Language, vol. 23, pp. 311-331, July 2009.

[15] A. Çıltık, and T. Güngör, "Time-efficient spam e-mail filtering using n-gram models," Pattern Recognition Letters, vol. 29, pp. 19-33, Jan. 2008.

[16] Ö. Özyurt, and C. Köse, "Chat mining: Automatically determination of chat conversations," Expert Systems with Applications, vol. 37, pp. 8705-8710, Dec. 2010.

[17] L. Özgür, T. Güngör, and F. Gürgen, "Adaptive anti-spam filtering for agglutinative languages: a special case for Turkish," Pattern Recognition Letters, vol. 25, pp. 1819-1831, Dec. 2004.

[18] E. Alparslan, A. Karahoca, and H. Bahşi, "Classification of confidential documents by using adaptive neurofuzzy inference systems," Procedia Computer Science, vol. 3, pp. 1412-1417, 2011.

[19] A. K. Uysal, and S. Gunal, "The impact of preprocessing on text classification," Information Processing and Management, vol. 50, pp. 104-112, Jan. 2014.

[20] F. Türkoğlu, B. Diri, and M. F. Amasyalı, Author Attribution of Turkish Texts by Feature Mining, D. –S. Huang, L. Heutte, M. Loog, Ed. Berlin, Germany: Springer-Verlag, 2007.

[21] D. M. Christopher, and S. Hinrich, Foundations of statistical natural language processing, 4th ed., Cambridge, Massachusetts: MIT Press, 2001.

[22] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," Journal of Documentation, vol. 60, pp. 493-502, 2004.

[23] K. S. Jones, "IDF term weighting and IR research lessons," Journal of Documentation, vol. 60, pp. 521-523, 2004.

[24] J. L. Solka, "Text Data Mining: Theory and Methods," Statistics Surveys, vol. 2, pp. 94-112, 2008.

[25] J. -S. Xu, and Z. -O. Wang, "Tcblsa: A New Method Of Text Clustering," in Proc. Second International Conference on Machine Learning and Cybernetics, 2003, p. 63.

[26] W. Zhang, T. Yoshida, and X. Tang, "A comparative study of TF*IDF, LSI and multi-words for text classification," Expert Systems with Applications, vol. 38, pp. 2758-2565, Mar. 2011.

[27] Y. Yang, and J. O. Pedersen, "Comparative Study on Feature Selection in Text Categorization," in Proc. ICML-97, 1997, p. 412.

[28] V. N. Vapnik, The Nature of Statistical Learning Theory, 2nd ed., M. Jordan, S. L. Lauritzen, J. F. Lawless, V. Nair, Ed. New York, USA: Springer-Verlag, 2000.

[29] T. Joachims, "Text categorization with support vector machines: Learning with many relevant feature," in Proc. ECML-98, 1998, p. 137.

[30] E. Leopold, and J. Kindermann, "Text Categorization with Support Vector Machines. How to Represent Texts in Input Space?," Machine Learning, vol. 46, pp. 423-444, 2002.

[31] A. Wang, W. Yuan, J. Liu, Z. Yu, and H. Li, "A novel pattern recognition algorithm: Combining ART network with SVM to reconstruct a multi-class classifier," Computers & Mathematics with Applications, vol. 57, pp. 1908-1914, June 2009.

[32] S. Dumais, J. Platt, D. Heckerman, and M. Sahami, "Inductive learning algorithms and representations for text categorization," in Proc. CIKM '98, 1998, p. 148.

[33] (2014) Fatih University Computer Engineering Website. [Online]. Available: http://nlp.ceng.fatih.edu.tr/blog/tr/?p=31/

[34] (2014) Zemberek Website. [Online]. Available: https://code.google.com/p/zemberek/

[35] M. Radovanovic, and M. Ivanovic, "Text Mining: Approaches and Applications," Novi Sad J. Math., vol. 38, pp. 227-234, 2008.

[36] (2014) Kemik Website. [Online]. Available: http://www.kemik.yildiz.edu.tr/?id=28/

[37] E. Alpaydın, Introduction to Machine Learning, 2nd ed., London, England: MIT Press, 2010.