Examining the Impact of Feature Selection Methods on Text Classification

Mehmet Fatih KARACA
Department of Computer Technologies
Gaziosmanpasa University
Tokat, TURKEY

Safak BAYIR
Department of Computer Engineering
Karabuk University
Karabuk, TURKEY

***Abstract*—Feature selection that aims to determine and select the distinctive terms representing a best document is one of the most important steps of classification. With the feature selection, dimension of document vectors are reduced and consequently duration of the process is shortened. In this study, feature selection methods were studied in terms of dimension reduction rates, classification success rates, and dimension reduction-classification success relation. As classifiers, kNN (k-Nearest Neighbors) and SVM (Support Vector Machines) were used. 5 standard (Odds Ratio-OR, Mutual Information-MI, Information Gain-IG, Chi-Square-CHI and Document Frequency-DF), 2 combined (Union of Feature Selections-UFS and Correlation of Union of Feature Selections-CUFS) and 1 new (Sum of Term Frequency-STF) feature selection methods were tested. The application was performed by selecting 100 to 1000 terms (with an increment of 100 terms) from each class. It was seen that kNN produces much better results than SVM. STF was found out to be the most successful feature selection considering the average values in both datasets. It was also found out that CUFS, a combined model, is the one that reduces the dimension the most, accordingly, it was seen that CUFS classify the documents more successfully with less terms and in short period compared to many of the standard methods.**

Keywords—Feature selection; text classification; text mining; k-Nearest Neighbors; support vector machines

I. Introduction

Text based data amounts reached enormous sizes on the web as a result of increasing number of computers, tablets and smart phones and their widespread use. This fact resulting from widespread use of technology caused changes in people’s habits. One of the instances of this is the topic of this paper which is news portals.

Busy schedule at work and the desire to catch up with frequently changing state agendas increased the use and significance of news portals. Columnists are one of the features that readers follow mostly on a news portal. A columnist may refer to various topics, in other words, more than one topic, write about a topic outside his area of interest and even title of his article may not be consistent with the content, being incoherent. Therefore, classification of an article in terms of its topic is important in order to give information about its content to readers.

Since articles contain unstructured data, it is not possible to analyze articles directly through data mining techniques. Text mining provides an opportunity to apply data mining techniques by converting unstructured text based data into structured form. Text mining is used to extract the unknown and useful information with the analysis of unstructured documents for specific purposes [1]. On the other hand, data mining extracts concealed and potentially useful information from available data [2]. It is necessary to filter, govern and classify data for people to get a quick access to information [1]. Text classification refers to the assignment of texts to pre-determined categories. Prior to computer systems, classification was done manually. This process was not only slow and expensive but also inconsistent. That the processes are done via computers decreases those problems to a great extent.

In text classification studies, it is seen that preprocessing, feature selection methods, term weighting and classification algorithms are taken into consideration. In this study, the feature selection methods, which both decrease the duration of the process and provide opportunity to make successful classification, were taken into account. Standard methods were applied either directly or variously, besides, a new feature selection method was tested. Turkish corpus consisted of columns which were formed for this study and English corpus titled as 20Newsgroups were used as datasets.

The organization of the paper is as follows: Methodology of the study is given in Section 2, experimental results are provided and discussed in Section 3 and the conclusion part is included in Section 4.

A. Related Work

Classification is one of the most researched and studied text mining subjects. Text mining which does not only consist of classification, also includes unstructured data analysis such as topic/author detection, spam e-mail filtering, table/report analysis, document summarization, and question/answering systems. Unstructured data passes through a series of processes while it is being converted into structured form; preprocessing, feature selection, term weighting and finally obtaining document vectors respectively. One or several of these steps were dealt with together in studies. Reuters-21578 [3]-[10] and 20Newsgroups [5], [6] datasets, consisting of English text content, are widely used to provide a general evaluation related to applied methods. Datasets which are composed of different sources and languages such as e-mail [4], SMS [4], news text [11], [12], technical paper [9], medical journals [13] and chemical web pages [10] are used to reveal the effect of classification methods on the other languages. Datasets containing Turkish documents are limited in number and they are not regarded as standard datasets yet. Some of them are as follows; 6-class 2 imbalanced datasets formed with news obtained from RSS source [11], and 5, 6 and 9-class 3 balanced datasets formed with columns and news [12]. Since there is not a standard dataset consisting of Turkish content, the evaluation of effects of the techniques on Turkish content cannot be done.

Feature selection is the process of determining the terms to be used in classification. It is not only dimension reduction of document vectors both also ensures better results [7], [14] and decreases process time. Feature selection is applied almost in all text classification studies. Moreover, there are studies in which only feature selection techniques are evaluated. Document frequency, mutual information, information gain and chi-square are the most widely used feature selection methods [5], [6], [13]-[15]. Besides, studies displayed that hybrid models of filter and wrapper are applied and better results are produced [5]. Liu et al. [9] used feature selection methods for term weighting in their studies. Furthermore, a feature selection may not have the same effect on all classification algorithms; a feature selection producing the best results for an algorithm may not necessarily produce the same results for another algorithm [12].

Classification is the process of assigning documents to predefined classes. Classification process is carried out through computing the relationship between test document and training document vectors and their classes with methods such as kNN, Support Vector Machines, NaïveBayes and Artificial Neural Networks. kNN becomes one of the most preferred algorithms as a result of having uncomplicated formulae which are to be used in calculation operations and similar reasons. SVM aiming to form n-dimension hyper plane in order to separate classes, also, is one of the most preferred classifiers. In their study, Karaca et al. [16] studied similarity calculation techniques for kNN. It was found out that the best techniques differ depending on whether feature selection is applied or not. In their study, Yang and Pedersen [13] applied kNN and Cosine together, and stated that information gain and chi-square are more effective than document frequency, mutual information and term strength. It was reported that document frequency is used instead of information gain and chi-square, because carrying out the computation with these two measures is too expensive compared to document frequency. Uysal and Gunal [4] used SVM as classifier, Zemberek and Porter for stemming and chi-square for feature selection in their studies which mainly focus on the effects of preprocessing upon classification. It was reported that there is not any successful preprocessing method in each domain and language. Günal [5] applied Decision Tree and SVM as classifiers as well as mutual information, chi-square, information gain as feature selections, and tested their hybrid models, and stated that hybrid models produce better results. Liang et al. [10] preferred dictionary-based approach in their studies. In the study by Sanwaliya et al. [8], Decision Tree, Rocchio, NaiveBayes, NaiveBayes-kNN and kNN were used as classifiers, *k* value which was increased from 30 to 90 with an increment of 10 was tested and the best result was obtained when *k* is 50.

B. *Purpose of the Study*

This study aims to analyze the effects of feature selection methods on text classification. Besides 5 standard (Odds Ratio-OR, Mutual Information-MI, Information Gain-IG, Chi-Square-CHI and Document Frequency-DF), 2 combined (Union of Feature Selections - UFS and Correlation of Union of Feature Selections - CUFS) and 1 new (Sum of Term Frequency - STF) feature selection methods were tested by utilizing 2 datasets. With this study, it is aimed to form a perspective on feature selection methods regarding examination of the following issues:

- Effects of standard feature selection methods on dimension reduction and classification success.
- Effects of combined feature selection methods on dimension reduction and classification success.
- Effects of new feature selection method on dimension reduction and classification success.
- Dimension reduction-classification success relationship.

II. Methodology

Process steps that must be followed in order to classify the documents are shown in Fig. 1. Firstly, document must be input into the system and undergone the preprocess, then, feature selection (if it is a training document) and term weighting must be applied respectively, later, document vectors must be obtained and finally, classification must be carried out.

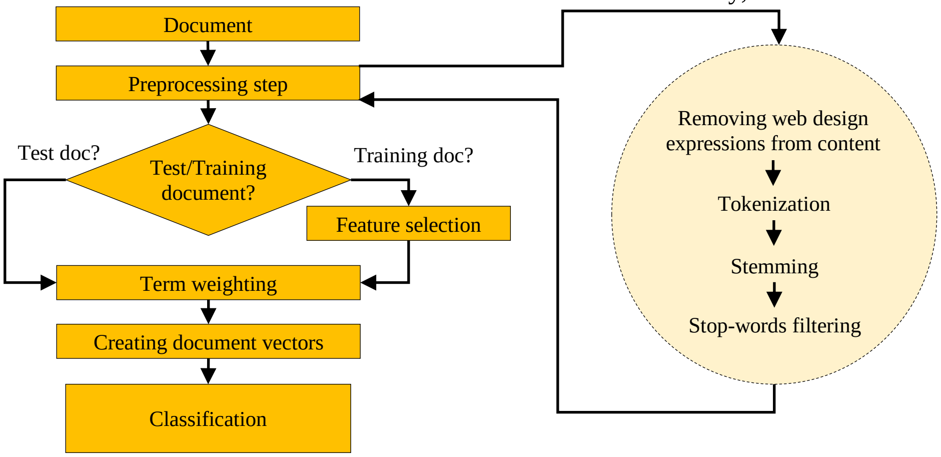


Fig. 1. Classification steps of documents.

TABLE I. SAMPLE NUMBERS OF DATASETS

ColumnDataset				20Newsgroups			
Class Label	**#Training**	**#Testing**	**#Total**	**Class Label**	**#Training**	**#Testing**	**#Total**
Economy (*Ekonomi*)	675	225	**900**	talk.politics.misc	675	225	**900**
Sport (*Spor*)	675	225	**900**	rec.sport.hockey	675	225	**900**
Health (*Sağlık*)	675	225	**900**	sci.med	675	225	**900**
Education (*Eğitim*)	675	225	**900**	sci.space	675	225	**900**
Life (*Yaşam*)	675	225	**900**	sci.crypt	675	225	**900**
#Total	**3375**	**1125**	**4500**	**#Total**	**3375**	**1125**	**4500**

All the processes except feature selection are applied to both training and test documents. Besides, it must be specified that preprocessing is made up of various sub processes that can differ depending on type and language of the document.

A. *Datasets*

Two datasets were used in this study: ColumnDataset and 20Newsgroups. ColumnDataset consists of columns and contains document content in Turkish. This dataset was created by extracting articles of a total of 35 columnists from their official news sites on a daily basis between the dates of 06.09.2006 and 02.12.2014 with a real-time crawler within the software developed. The other dataset titled as 20Newsgroups includes English texts and it is commonly used in text mining studies [17].

Information regarding the datasets used within the study was given in Table I. As seen, in these balanced two datasets with five classes each, there is a total of 4500 documents including 675 training and 225 testing samples in each class.

B. *Preprocessing*

Preprocessing is the first step to convert unstructured data into structured form. Preprocessing, one of the most important steps of text mining studies, may change depending on the document language, type and the source it is obtained. However, in order to obtain a pure document, the following preprocessing steps are carried out; tokenization, stemming and stop-words filtering.

Tokenization is the process of breaking the documents into words, called token. After this step, processes are performed as word-based. The stemming method which is going to be applied to a Turkish or English document is not the same, since grammar rules of these languages are different. For stemming process, Zemberek [4] is generally preferred in Turkish documents while Porter [4], [5], [8], [10] is used for English documents. In this study, Zemberek [18], an open source Turkish Natural Language Processing Library, was used for ColumnDataset documents, and Porter [19] was used for 20Newsgroups documents. Stop-words, which occur frequently in documents, do not provide any insight about the text within the document and also do not have meaning on their own [4] were determined and removed from the documents.

C. *Feature Selection*

Feature selection aims to determine and select the distinctive terms representing a document best [6]. One of the biggest obstacles in text classification is high-dimensional feature space [13]. Through feature selection, terms to be used in classification process are determined, dimension of feature space is reduced and thus duration of the process is shortened [9], [20]. The better the terms that represent the document are chosen, the higher the classification success becomes. Moreover, studies reveal that better results are obtained when feature selection is applied [7], [14].

In this study, information regarding eight feature selection methods in total including five standard methods can be seen in Table II. Feature selection, in this study, was applied as follows; generic term pools for each class were created out of each class containing 100, 200, 300, 400, 500, 600, 700, 800, 900 or 1000 terms. Then, feature space was created by combining the generic term pools.

For instance, document frequency values of terms for each class were calculated and 100 terms with the highest values from each class were selected and combined. As a result of this, DF (100) feature vector was created. OR, MI, IG, CHI and DF are the standard feature selection methods used in this study. In most studies [5], [9], [13], [15], [21], [22] one or several of these selection methods are used. The values of each term for each class are computed separately in OR, MI, IG and CHI methods and the terms are determined considering these values. On the other hand, in DF method, determination process of terms is performed via a SQL query without requiring calculation of terms. UFS and CUFS are the combined models of standard methods. UFS is a combination of standard feature selections without any criterion. For instance, UFS (100) was created by the union of terms detected by OR (100), MI (100), IG (100), CHI (100), and DF (100). Therefore, vector dimension obtained with UFS is relatively much higher than standard methods.

The correlation values between each term (v_t) obtained from UFS and classes (v_c) were calculated and the absolute values of these values were sorted in a descending order. Then, CUFS was created as a result of selecting specific number of terms (e.g. 100) that have the highest values among them. For instance, correlation values of the terms resulting from UFS (100) were calculated; CUFS (100) was formed by choosing the first 100 terms with the highest values. Minimum vector dimension was reached through this method.

STF is similar to DF method but utilizes a new approach. DF deals with the number of documents where a term occurs while STF deals with the frequency of term occurrences across the documents.

TABLE II. FEATURE SELECTION METHODS

Type	Name	Label	Number of Selected Terms	Formula*
Standard	Odds Ratio	OR	Top 100 to 1000 terms (with an increment of 100 terms) from each class	$= \log\left(\frac{AD}{BC}\right)$
	Mutual Information	MI		$= \log\left[\frac{AN}{(A+B)(A+C)}\right]$
	Information Gain	IG		$= -\frac{A+C}{N}\log\left(\frac{A+C}{N}\right) + \frac{A}{N}\log\left(\frac{A}{A+B}\right) + \frac{C}{N}\log\left(\frac{C}{C+D}\right)$
	Chi-Square	CHI		$= \frac{N(AD-BC)^2}{(A+C)(B+D)(A+B)(C+D)}$
	Document Frequency	DF		$SQL\ Query\ (COUNT)$
Combined	Union of Feature Selections	UFS	Terms from union of standard methods	-
	Correlation of Union of Feature Selections	CUFS	Top 100 to 1000 terms (with an increment of 100 terms) from UFS (considering correlation)	$= \left\lvert \frac{N\sum v_t v_c - \sum v_t \sum v_c}{\left\{\left[N\sum v_t^2 - \left(\sum v_t\right)^2\right]\left[N\sum v_c^2 - \left(\sum v_c\right)^2\right]\right\}^{1/2}} \right\rvert$
New	Sum of Term Frequency	STF	Top 100 to 1000 terms (with an increment of 100 terms) from each class	$SQL\ Query\ (SUM)$

*A: Number of documents belonging to class k that the term occurs.
B: Number of documents not belonging to class k that the term occurs.
C: Number of documents belonging to class k that the term does not occur.
D: Number of documents not belong to class k that the term does not occur.
N: Training documents number.
v_t: variable1 (the term), v_c: variable2 (classes).

With STF, term occurrence in a document becomes more significant, ensuring that dominant terms in a given class are emphasized. For instance, the word “game” is a sports term and occurs more frequently in sports documents. If the term is evaluated once as it is in DF, effect of the term for sports class will be decreased. Process of term determination in STF is carried out with SQL query as in DF.

D. Term Weighting

Terms to be used in classification are determined through feature selection. These terms are weighted in a various ways depending on the number of occurrence within the documents. Then, weighted terms are united and document vector is created. Term weighting can be referred as value or impact of a term in document [23].

In this study, binary weighting, one of the simplest weightings that deals with presence/absence of a term in a document was used. The process of converting unstructured documents into structured form was completed with the numerical expression of terms in document as a result of weighting. This process starts with preprocessing and ends with term weighting and formation of document vectors [12].

E. Classification

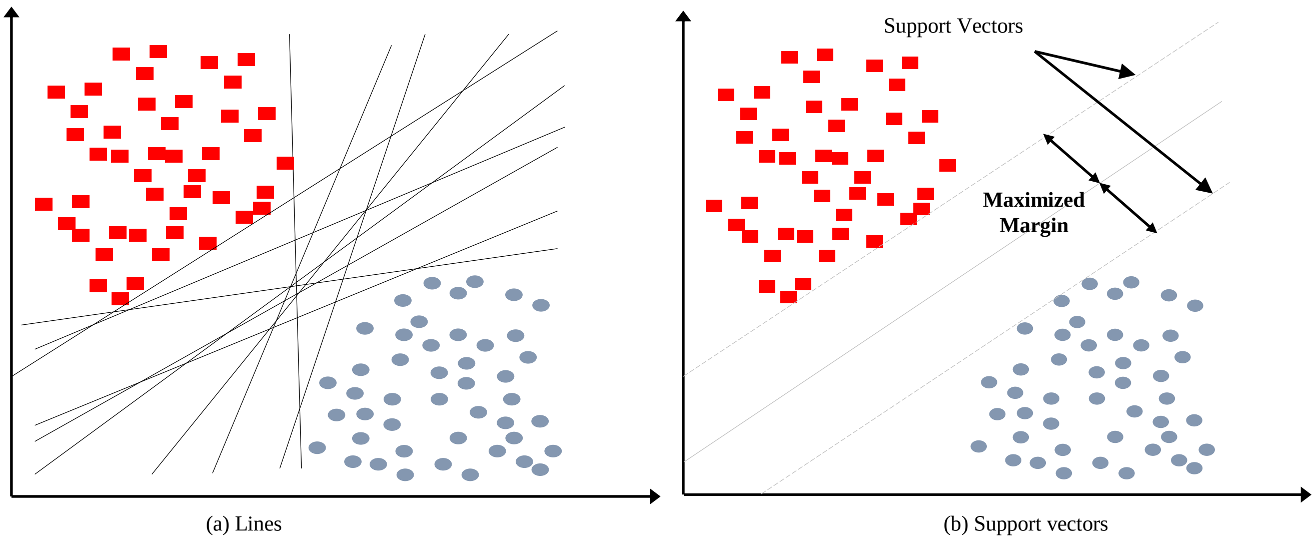
Text classification is the process of assigning natural language texts to a predefined classes with a classification algorithm. In this study, kNN and SVM were used as classifiers. The advantage of kNN is that it does not require training of the system; however, SVM does.

1) kNN Classifier: Primarily in kNN, similarity between test and training document vectors are calculated with a variety of techniques such as Cosine similarity, Euclidean distance and inner product. Then, similarity values are sorted, and class with the highest frequency within k document is assigned as the class of test document [24].

In this study, the classification is carried out with Cosine similarity, Euclidean distance, harmonic mean and inner product. However, since more successful results were obtained with Cosine similarity, only the results belonging to Cosine similarity were presented. Equation for computing the similarity between X and Y vectors by using with Cosine similarity was given in (1). k value was determined as 7 (flexible). As of the 7th document, when any document with the same similarity value with the 7th document and belonging to a different class was determined, k value was increased at that rate (number). For instance, when 8th (health) and 9th (education) vectors have the same similarity value with the 7th (sports) vector and different classes, k value is increased to 9.

$$\cos(\theta) = \frac{XY}{\|X\|\|Y\|} \tag{1}$$

2) SVM Classifier: Support Vector Networks, later named as Support Vector Machines, was introduced by Vladimir Vapnik and applied for two-group classification in 1995 [25].



(a) Lines (b) Support vectors

Fig. 2. Generating hyper planes for separating groups.

Although there are many lines in order to separate the groups (Fig. 2(a)), the main aim of SVM is to achieve the maximum possible margin and optimal hyper plane (Fig. 2(b)) where the best classification will be realized [6]. In this study, linear kernel function which is claimed to generate better results for multi-class classification than other kernel functions [12] was preferred.

F. Success Measures

In this study, MacroF1 (F-measure) shown in (2) was applied to determine the classification success. At first, precision (p_k) and recall (r_k) values of the classes are calculated. tp_k (true positive) denotes the number of documents belonging and assigned to class k, fp_k (false positive) denotes the number of documents not belonging but assigned to class k, fn_k (false negative) denotes the number of documents belonging but not assigned to class k, n_k denotes the number of classes (Fig. 3).

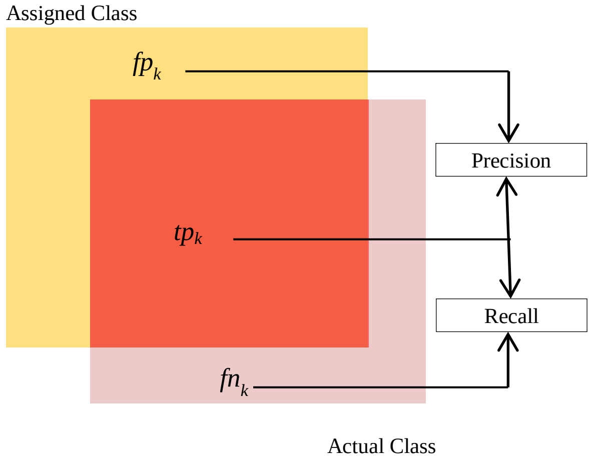


Fig. 3. Precision and recall schema.

Since equal number of classes exists in both two datasets and also equal number of documents exists in each class, primarily F_k value is calculated for each class, then, averages of F_k were calculated, and MacroF1 was computed as % multiplying by 100. In this study, it was seen that each test document is essentially assigned to a class.

$$p_k = \frac{tp_k}{tp_k + fp_k}$$

$$r_k = \frac{tp_k}{tp_k + fn_k}$$

$$F_k = 2\frac{p_k r_k}{tp_k + r_k} \qquad (2)$$

$$MacroF1 = 100\frac{\sum_1^{n_k} F_k}{n_k}$$

III. Experimental Results

In this study, feature selection methods were analyzed from different perspectives. Classification was carried out by using 2 combined and 1 new feature selection methods as well as 5 standard feature selections. Two datasets were used to evaluate the effects of the methods. The first dataset is ColumnDataset which consists of columns from newsportals and includes Turkish texts, and the other dataset is 20Newsgroups including English documents. Number of classes and documents in the classes in both datasets are equal.

Results were evaluated in terms of dimension reduction ratio of feature selection methods and their effects on classification success. Results were provided separately for either dataset as an average of kNN and SVM. When the figures providing results are taken into consideration, it is seen that almost parallel results are obtained in both datasets.

Number of unique terms in ColumnDataset is 11528, number of total usage of terms is 1377787, average number of terms in a document is 306. Number of unique terms in 20Nwesgroups is 28458, number of total usage of terms is 620004, average number of terms in a document is 138. These figures show that documents in ColumnDataset are relatively longer than those in 20Newsgroups in terms of the total number of terms.

It can be said that Zemberek performs better than Porter in detecting the root/stem of the words. This situation can be explained with three examples provided by Porter. First, the root of the letter series of “lbtlk” which was randomly entered was determined as “lbtlk”, but there is not a such word in English. Second, although the original root of the word “agree” is “agree”, the root of the word “agree” was determined as “agre”, therefore, original root could not be determined. Third, and the most important example is that the root of the word “focus” was determined as “focu” and the root of the word “focusing” was determined as “focus”. But the root of these two words is “focus”. Therefore, the words having the same root will be evaluated as if they are different words. One of the possible reasons of having low classification success in 20Newsgroups compared to ColumnDataset can be this issue.

Dimension reduction ratio of feature selection methods are presented in Fig. 4 and 5. According to feature vectors, when 1000 terms were selected from each class, 512 common terms were detected in ColumnDataset while no common word was found in 20Newsgroups; it can be said that this situation was occurred as a result of the difference between Zemberek and Porter.

The maximum dimension reduction ratio in both datasets was achieved with CUFS feature selection method depending on correlation. CUFS achieved 90.78% to 99.08% success in ColumnDataset and 95.99% to 99.60% in 20Newsgroups, which were quite high success rates of dimension reduction. When the average dimension reduction rates were taken into consideration, it was found out that DF, focusing on the number of documents in which terms occur across, reached a high ratio of dimension reduction which was not as much as CUFS does, though. A new approach, STF, focusing on total number of terms occurrences across documents, achieved a very close dimension reduction ratio to that of DF. The minimum dimension reduction ratio was observed in UFS which is a combination of standard methods.

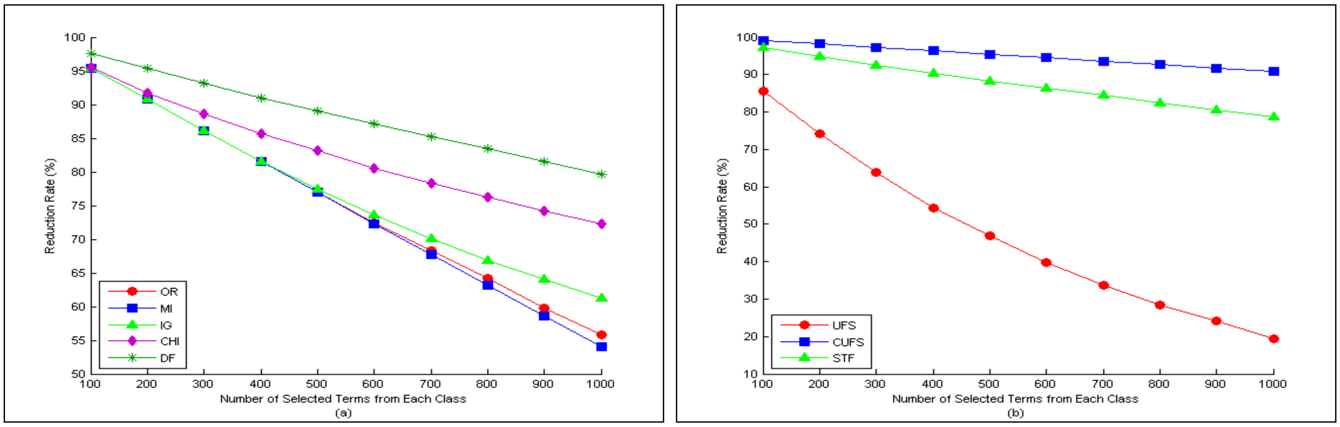


Fig. 4. Reduction rate of feature vector dimension on ColumnDataset: (a) standard methods; (b) combined & new methods.

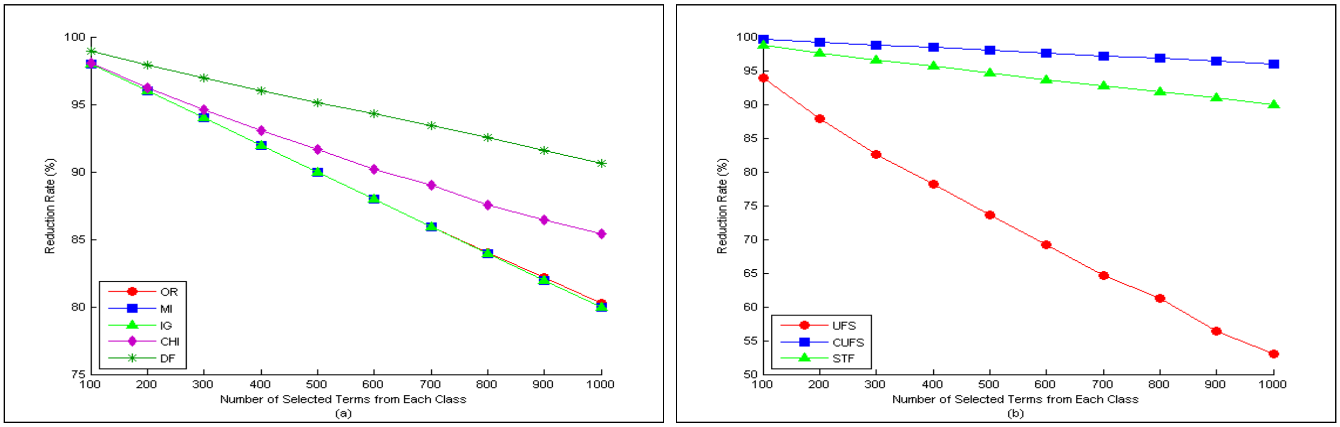


Fig. 5. Reduction rate of feature vector dimension on 20Newsgroups: (a) standard methods; (b) combined & new methods.

TABLE III. MACROF1 VALUES ON COLUMN DATASET

	kNN								**SVM**							
	OR	**MI**	**IG**	**CHI**	**DF**	**UFS**	**CUFS**	**STF**	**OR**	**MI**	**IG**	**CHI**	**DF**	**UFS**	**CUFS**	**STF**
100	98.25	51.60	52.38	99.16	97.56	99.33	97.32	98.49	97.69	55.43	55.20	98.67	97.96	97.96	93.03	98.76
200	98.53	70.55	71.87	99.29	98.44	98.89	98.18	98.75	97.68	65.94	64.99	98.93	98.67	82.85	94.52	98.58
300	99.15	80.00	80.86	99.24	98.53	98.84	98.58	98.80	96.18	72.63	73.30	98.22	98.76	68.85	95.64	98.67
400	99.02	85.37	83.99	99.02	98.71	99.02	98.36	98.98	94.93	75.84	78.40	97.61	98.23	91.74	96.10	98.76
500	98.84	92.64	98.00	99.24	99.11	99.15	98.71	98.93	86.96	82.33	90.15	96.98	98.23	93.99	96.63	97.79
600	99.11	94.05	98.66	99.02	99.24	99.02	98.71	98.98	74.95	80.64	88.86	95.46	97.70	96.22	96.36	97.52
700	98.84	94.48	99.06	99.15	99.29	99.20	98.80	99.11	60.95	70.00	85.01	93.93	96.72	96.66	95.90	96.99
800	98.57	95.67	99.15	99.06	99.29	99.28	98.62	99.02	77.52	68.05	80.07	90.27	96.19	97.37	95.99	96.10
900	98.44	95.82	99.06	99.24	99.06	99.06	98.57	99.02	87.37	59.72	75.31	85.91	95.85	97.21	96.34	95.75
1000	98.75	95.17	99.20	99.15	99.02	99.20	98.80	99.15	92.63	71.48	68.97	78.62	94.53	97.64	96.35	94.32

TABLE IV. MACROF1 VALUES ON 20NEWSGROUPS

	kNN								**SVM**							
	OR	**MI**	**IG**	**CHI**	**DF**	**UFS**	**CUFS**	**STF**	**OR**	**MI**	**IG**	**CHI**	**DF**	**UFS**	**CUFS**	**STF**
100	89.35	45.16	48.10	92.25	86.88	90.14	74.81	87.80	84.13	43.38	43.60	86.96	83.26	82.30	76.07	83.93
200	90.84	58.96	50.60	93.90	89.15	93.07	78.45	90.90	82.88	53.14	47.44	86.75	83.70	67.32	79.70	83.56
300	92.00	63.55	63.62	93.39	91.59	93.19	82.15	92.32	80.47	57.32	56.67	79.68	83.71	46.06	79.79	83.46
400	91.89	67.55	68.44	94.49	92.11	93.23	83.87	92.21	72.60	59.26	60.10	77.44	82.92	61.01	78.54	83.22
500	92.01	70.06	70.54	93.95	91.94	93.45	85.73	92.39	63.86	59.23	58.88	71.32	81.17	73.72	79.47	81.84
600	93.07	72.47	71.87	94.96	92.76	94.13	85.92	93.13	49.93	60.79	59.84	64.07	79.80	77.92	80.29	78.85
700	93.00	73.61	72.80	94.27	92.60	93.59	86.03	93.04	44.06	61.14	60.85	57.86	76.33	80.78	79.82	75.99
800	93.56	74.88	75.65	93.48	92.33	94.38	87.09	93.37	49.94	62.53	63.61	48.96	74.44	81.51	79.43	73.32
900	93.77	75.50	74.82	94.05	92.44	94.09	87.82	93.10	54.23	64.58	64.67	44.58	72.57	83.97	79.08	70.08
1000	94.02	78.10	76.84	93.23	92.52	94.25	88.02	93.68	56.74	64.23	63.94	47.83	67.62	83.93	78.15	64.14

All classification results related to all techniques applied within the scope of this study were provided in Tables III and IV. 80 classifications by kNN and SVM, 160 classifications in total, were performed in both ColumnDataset and 20Newsgroups. It was seen that kNN generate better results than SVM, and effect of Cosine on kNN's success was obvious. Besides, Cosine calculates the similarity with an approach regarding the terms existing in both training and testing documents and also the norms of document vectors.

Classification success rates of standard, combined and new methods are seen in Fig. 6 and 7. When Fig. 6 is examined, it is seen that STF and DF are not severely affected by the number of terms in comparison with other methods. The most successful classification was performed with CHI (200), a standard method and with STF (400), a new method. Although the feature vector dimension of CUFS is low, it has a high level of classification success. According to Fig. 7, where results of 20Newsgroups are shown, the most successful results were produced with the standard method of CHI (200) and with the combined method of UFS (1000). DF and STF which use similar techniques were not affected from increase in number of selected terms when compared to other methods, this situation can be the result of their more effective feature selection implementation. Moreover, the effect of increase in number of selected terms was found out as low in ColumnDataset when compared to 20Newsgroups, this situation can be the cause of Zemberek.

Average values related to feature selection methods and the number of terms selected from each class is seen in Fig. 8. According to Fig. 8(a), STF produces the best results in both datasets. Although DF produces close results to STF, it was found out that STF performed classification much more successfully than DF. When Fig. 8(b) is taken into consideration, it was seen that the best results in both datasets are obtained when 500 terms are chosen from each class. That STF produce better results compared to DF can be the result of considering the frequency of term occurrences across the documents instead of the number of documents where a term occurs, this reveals that this approach provides more accurate results. Furthermore, dominant terms within their class become more significant with STF compared to DF.

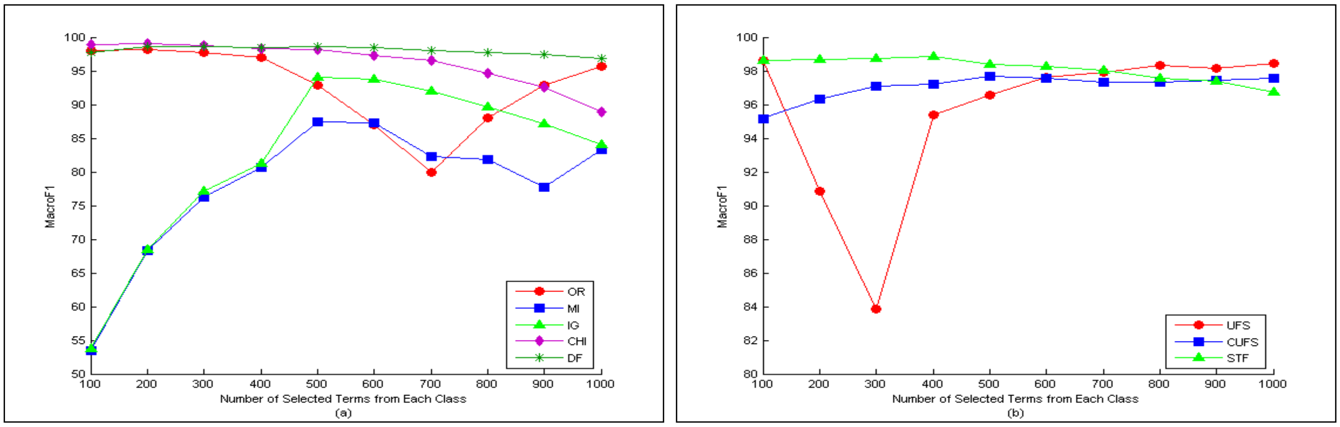


Fig. 6. Success measures on ColumnDataset: (a) standard methods; (b) combined & new methods.

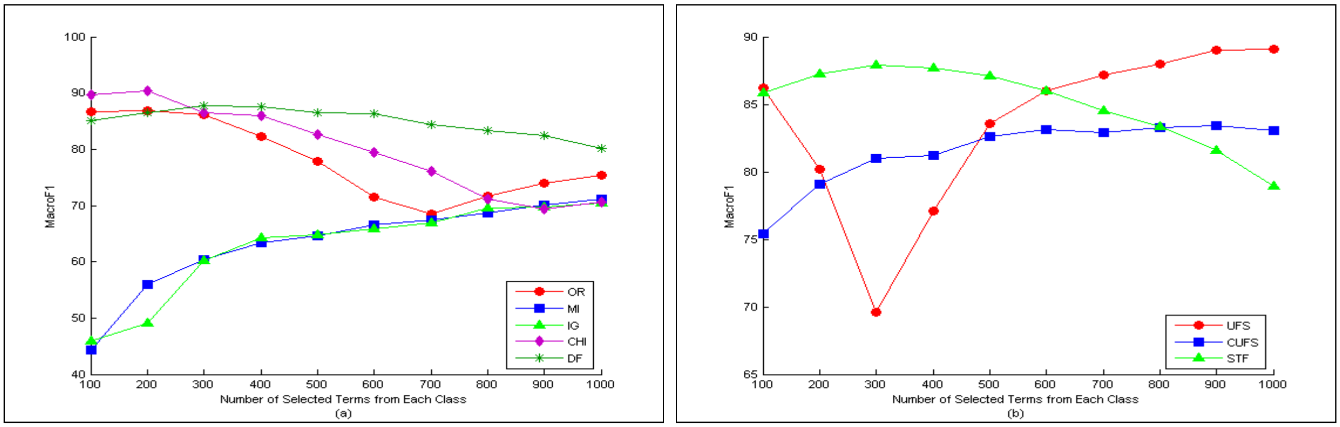


Fig. 7. Success measures on 20Newsgroups: (a) standard methods; (b) combined & new methods.

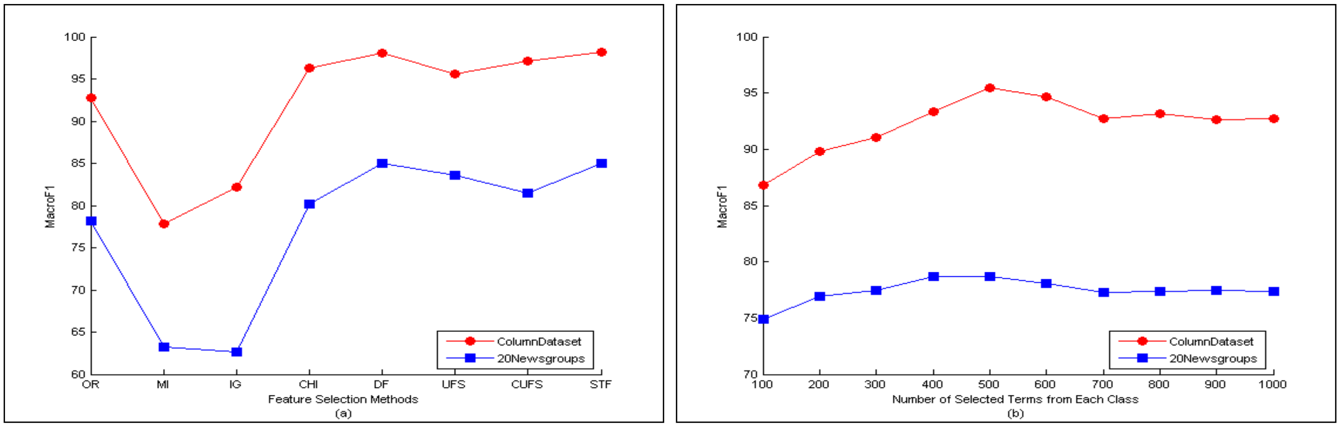


Fig. 8. Average success measures: (a) feature selection methods; (b) number of selected terms from each class.

IV. Conclusion

In this study, feature selection methods, highly significant subject in text classification, were studied. These methods were tested by selecting 100 to 1000 terms (with an increment of 100 terms) from each class. The results obtained can be summarized as follows:

- DF was the standard method in which the most dimension reduction and the best classification were realized.
- CUFS was the combined method which reduced the dimension most.
- There was a 2% difference between CUFS and UFS which is the other combined method in classification success.
- STF, a new method, provided dimension reduction with the similar values to CUFS.
- When all the feature selection methods are evaluated together, it was found out that the maximum dimension

reduction is obtained with CUFS and the most successful classification is obtained with STF.

- CUFS, STF, CHI and DF methods were affected less from the increase in number of terms compared to other methods.
- It can be said that STF was primarily preferred as a result of having most successful results despite reducing the rate of feature vector dimension seriously.
- It was observed that kNN was predominantly successful compared to SVM.

When the results obtained with two datasets were taken into consideration in terms of trends, it was seen that the graphics display similarity in a parallel fashion. This situation revealed that the methods used in the study were utilized appropriately in order to make a general evaluation. Despite the fact that the trends of the graphics displayed similarity in two datasets, it was found out that MacroF1 values were different. Besides, it was observed that Zemberek (used for Turkish) is more successful than Porter (used for English) in the detection process of the root/stem of the words. It can be said that this situation has an effect in obtaining more successful results with the dataset including Turkish content.

In the future studies, preprocessing, term weighting and classification algorithms which are the other factors affecting text classification success can be examined.

References

[1] A. Visa, Technology of text mining, Lecture Notes in Computer Sciences, vol. 2123, pp. 1-11, 2001.

[2] W. J. Frawley, G. Piatetsky-Shapiro and C. J. Matheus, “Knowledge discovery in databases: An overview,” AI Magazine, vol. 13, no. 3, pp. 57-70, 1992.

[3] F. Colace, M. D. Santo, L. Greco and P. Napoletano, “Text classification using a few labeled examples,” Computers in Human Behavior, vol. 30, pp. 689-697, 2014.

[4] A. K. Uysal and S. Gunal, “The impact of preprocessing on text classification,” Information Processing and Management, vol. 50, no. 1, pp. 104-112, 2014.

[5] S. Günal, “Hybrid feature selection for text classification,” Turkish Journal of Electrical Engineering & Computer Sciences, vol. 20, no. 2, pp. 1296-1311, 2012.

[6] A. K. Uysal and S. Gunal, “A novel probabilistic feature selection method for text classification,” Knowledge-Based Systems, vol. 36, pp. 226-235, 2012.

[7] O. Durmaz and H. Ş. Bilge, “Metin sınıflandırmada boyut azaltmanın etkileri ve özellik seçimi,” 19th Conference on Signal Processing and Communications Applications (SIU 2011), Antalya, Turkey, pp. 21-24, 2011.

[8] A. Sanwaliya, K. Shanker and S. C. Misra, “Categorization of news articles: A model based on discriminative term extraction method,” 2nd International Conference on Advances in Databases, Knowledge, and Data Applications (DBKDA 2010), French Alps, France, pp. 145-154, 2010.

[9] Y. Liu, H. T. Loh and A. Sun, “Imbalanced text classification: A term weighting approach,” Expert Systems with Applications, vol. 36, no. 1, pp. 690-701, 2009.

[10] C. Y. Liang, L. Guo, Z-J. Xia, F-G. Nie, X-X. Li, L. Su and Z-Y. Yang, “Dictionary-based text categorization of chemical webpages,” Information Processing & Management, vol. 42, no. 4, pp. 1017-1029, 2006.

[11] Ç. Toraman, F. Can and S. Koçberber, “Developing a text categorization template for Turkish news portals,” International Symposium on Innovations in Intelligent Systems and Applications (INISTA 2011), İstanbul, Turkey, pp. 379-383, 2011.

[12] D. Torunoğlu, E. Çakırman, M. C. Ganiz, S. Akyokuş and M. Z. Gürbüz, “Analysis of preprocessing methods on classification of Turkish texts,” International Symposium on Innovations in Intelligent Systems and Applications (INISTA 2011), İstanbul, Turkey, pp. 112-117, 2011.

[13] Y. Yang and J. O. Pedersen, “A comparative study on feature selection in text categorization,” Proceedings of the 14th International Conference on Machine Learning (ICML '97), Nashville, Tennessee, USA, pp. 412-420, 1997.

[14] J. Chen, H. Huang, S. Tian and Y. Qu, “Feature selection for text classification with Naïve Bayes,” Expert Systems with Applications, vol. 36, no. 3, pp. 5432-5435, 2009.

[15] K. Aas and L. Eikvil, Text categorisation: A survey, Norwegian Computing Center, pp. 1-37, 1999.

[16] M. F. Karaca, M. Günel and A. A. Taştan, “Metin sınıflandırmada benzerlik hesaplama tekniklerinin değerlendirilmesi,” 17. Akademik Bilişim Konferansı, Eskişehir, Turkey, 2015.

[17] A. Asuncion and D. J. Newman, UCI Machine Learning Repository, University of California, Department of Information and Computer Science, Irvine, CA, 2007.

[18] Zemberek, http://code.google.com/p/zemberek (Accessed 2014).

[19] M. F. Porter, “An algorithm for suffix stripping,” Program, vol. 14, no. 3, pp. 130-137, 1980.

[20] C. M. Chen, H. M. Lee and C. C. Tan, “An intelligent web-page classifier with fair feature-subset selection,” Engineering Applications of Artificial Intelligence, vol. 19, no. 8, pp. 967-978, 2006.

[21] S.-B. Kim, K.-S. Han, H.-C. Rim, S. H. Myaeng, “Some effective techniques for Naive Bayes text classification”, IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 11, pp. 1457–1466, 2006.

[22] A. K. Uysal, “An improved global feature selection scheme for text classification,” Expert Systems with Applications, vol. 43, pp. 82-92, 2016.

[23] G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval,” Information Processing and Management, vol. 24, no. 3, pp. 513-523, 1998.

[24] J. Han and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, San Francisco, pp. 348-350, 2006.

[25] C. Cortes and V. Vapnik, “Support-vector networks,” Machine Learning, vol. 20, no. 3, pp. 273-297, 1995.