Engineering Applications of Artificial Intelligence 35 (2014) 38–53

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

A novel framework for termset selection and weighting in binary text classification

Dima Badawi, Hakan Altınçay*

Department of Computer Engineering, Eastern Mediterranean University, Famagusta, Northern Cyprus, Turkey

ARTICLE INFO

Article history:
Received 1 May 2013
Received in revised form
24 March 2014
Accepted 16 June 2014

Keywords:
Co-occurrence features
Termset selection
Termset weighting
Document representation
Text categorization

ABSTRACT

This study presents a new framework for termset selection and weighting. The proposed framework is based on employing the joint occurrence statistics of pairs of terms for termset selection and weighting. More specifically, each termset is evaluated by taking into account the simultaneous or individual occurrences of the terms within the termset. Based on the idea that the occurrence of one term but not the other may also convey valuable information for discrimination, the conventionally used term selection schemes are adapted to be employed for termset selection. Similarly, the weight of a selected termset is computed as a function of the terms that occur in the document under concern where a termset is assigned a nonzero weight if either or both of the terms appear in the document. This weight estimation scheme allows evaluation of the individual occurrences of the terms and their co-occurrences separately so as to compute the document-specific weight of each termset. The proposed termset-based representation is concatenated with the bag-of-words approach to construct the document vectors. Experiments conducted on three widely used datasets have verified the effectiveness of the proposed framework.

1. Introduction

Automatic text classification is one of the key tasks in various problems such as spam filtering in which the main aim is to get rid of unwanted emails, email foldering that aims to group the incoming messages into folders and sentiment classification where the main goal is to recognize whether a document expresses a positive or negative opinion. Because of this, text categorization has become an attractive research area for many researchers in the last two decades. One of the fundamental problems in text categorization is document representation. The conventional approach is the bag-of-words (BOW) (Sebastiani, 2002). In this representation, a subset of the terms that exist in the training collection is firstly selected after sorting them using a term selection measure such as χ^2, Gini index or information gain (IG) (Chen et al., 2009; Liu et al., 2009; Yang et al., 2012). Then, the document vectors are constructed using the frequencies and inverse document frequencies ($tf \times idf$) of the selected terms where the frequency of a term denotes the number of times it occurs in the document under concern. Alternatively, as a more simple method, binary representation is used where the feature value of a term is one if it appears in the document and zero otherwise. Experiments have shown that the feature value of a term, also known as its weight, can be more effectively calculated as the product of two factors, the term frequency and the collection frequency factors where the latter is used to take into account the discriminative abilities of different terms (Debole and Sebastiani, 2003).

In the BOW-based approach, the orders of words and their syntactic relations are not taken into account. As an extension to the BOW-based approach, the use of syntactic phrases and word sequences (n-grams) that are also known as statistical phrases is studied (Mladenic and Grobelnik, 1998; Lewis, 1992b). With the use of syntactic phrases, grammatical relations are also taken into consideration. Alternatively, n-grams which are generally defined as consecutive occurrences of pairs (bigrams) or triples of terms (trigrams) are employed to extract novel features (Caropreso et al., 2001; Tan et al., 2002; Bekkerman and Allan, 2004; Mladenic and Grobelnik, 1998). The main motivation for considering phrases is that a sequence of adjacent terms may be more discriminative than the individual terms in some cases. For instance, when considered individually, the terms “bill” and “gates” in the phrase “bill gates” may not be as informative as the phrase itself about the topic of the document (Bekkerman and Allan, 2004). Taking this into account, features representing phrases are defined where a phrase is said to occur if the corresponding sequence of adjacent terms appears in the document under concern. As another alternative, the use of termsets (or, compound features, itemsets)

* Corresponding author. Tel.: +90 392 6302842; fax: +90 392 3650711.
E-mail addresses: dima.badawi@emu.edu.tr (D. Badawi), hakan.altincay@emu.edu.tr (H. Altınçay).

http://dx.doi.org/10.1016/j.engappai.2014.06.012
0952-1976/

defined as the co-occurrences of terms having arbitrary order and position is also studied (Figueiredo et al., 2011; Tesar et al., 2006). In this approach, irrespective of their positions and order, if all terms appear, the corresponding termset is said to occur. Syntactic and statistical phrases are subsets of the set of all termsets. Since the number of termsets increases exponentially with the size of the vocabulary, termsets generally include pairs of terms but not triples. Experiments conducted on various datasets have shown that, when termsets or phrase-based features are concatenated with the BOW-based representation, better scores are generally achieved compared to the cases that exclude BOW and use only the termsets or phrases-based features (Lewis, 1992a; Boulis and Ostendorf, 2005).

As in the BOW-based approach, selection of a good subset of co-occurrence based features is important, and various criteria are utilized for this purpose. In his study on the use of syntactic phrases, Lewis (1992b) has argued that high dimensionality of the feature spaces, rare occurrence of distinct phrases and high redundancy due to synonymy are the major factors for achieving worse results compared to the BOW-based representation. Following his study, extensive work is carried out on selecting a good subset of co-occurring terms (Özgür and Güngör, 2010; Fürnkranz, 1998; Tan et al., 2002; Bekkerman and Allan, 2004). For instance, IG (Tan et al., 2002) and mutual information (MI) (Bekkerman and Allan, 2004) are used for selecting a subset of bigrams. Redundancy of features is a criterion that is considered for computing a discriminative set of features for text categorization (Baker and McCallum, 1998). This criterion is also used for selecting a good subset of bigrams. For instance, Boulis and Ostendorf (2005) argued that bigrams may not help improving the BOW representation when they are correlated with the features in the BOW-based representation, mainly due to the increased complexity especially when the training data is limited. They proposed a new measure to quantify the redundancy of a given bigram by considering the terms included in the bigram and reported improved accuracies on three different datasets. In a recent study, significant improvements compared to the BOW-based representation are achieved by applying pruning on both words and lexical dependencies (Özgür and Güngör, 2010). In fact, a weakness stated by Lewis is avoided by eliminating the rare words and the term dependencies with low occurrences. Figueiredo et al. (2011) underlined the importance of employing the most informative terms in termset generation. As a discrimination criterion, the number of classes in which the termsets appear is considered. Significantly better scores are achieved on four benchmark datasets by employing termsets of pairs of terms which are not restricted to be adjacent. The use of thresholds on the number of documents each phrase or termset appears in the training set is also considered in their selection (Figueiredo et al., 2011; Fürnkranz, 1998).

The studies mentioned above mainly aim at developing more intelligent schemes for selecting the best subset of phrases or termsets to be used together with BOW. However, in the case of BOW-based representation, term weighting is shown to be as important as selection and, various other measures such as relevance frequency and probability based scheme are proposed to replace the *idf* factor (Lan et al., 2009; Liu et al., 2009). Using these weighting schemes, it is also shown that significantly better performance scores can be achieved when compared to using binary or $tf \times idf$ based representation in the case of BOW. On the other hand, the termsets-based features are generally defined as binary where the feature value is computed as one if the corresponding termset appears (Figueiredo et al., 2011) and phrases-based features are defined as either binary or real-valued. In the case of real-valued features, only the frequencies are generally considered for their weighting.

In this study, a novel framework is proposed for selecting and weighting of termsets including non-adjacent pairs of terms. The idea is based on revising the definition of termset-based features. Consider a termset of two different terms. In the conventional representation, a termset is said to occur if both terms exist in the document. The proposed approach is based on utilizing the joint occurrence statistics of the terms for termset selection and weighting. More specifically, selecting and weighting termsets is performed by considering which term(s) occurred. The main motivation for this approach can be better explained by an example. Let us re-consider the "bill gates" example. If either of the terms is missing, the individual terms of the phrase are not as informative as the phrase itself as mentioned above. Hence, only the co-occurrence of these terms is deemed as valuable. However, there are other cases for which this phrase is not representative. For instance, consider the termset "tennis court". It can be argued that the occurrence of both terms supports the sports topic. But, different from the previous example, the occurrence of the first term without the second term also supports the same topic. Hence, it may be useful to assign large weights to the termset in both of these cases. The occurrence of the second term but not the first may also be statistically valuable. For instance, it may signify a different topic such as law. In other words, the term "court" may not be discriminative on its own since it appears in both sports and law related documents, but it becomes more informative when evaluated together with "tennis". It can be concluded that co-occurrence is not essential for a termset to represent valuable information. As a matter of fact, instead of focusing on only the co-occurrence of the terms, evaluation of all three possibilities in selecting and weighting termsets is promising. In this study, the joint occurrences of the individual terms within the termsets including two terms are investigated for their selection and weighting. The conventionally used selection and weighting schemes are adapted to employ this information. Experiments conducted on three widely used benchmark datasets have shown that the proposed scheme is remarkably superior to the baseline.

The rest of this paper is organized as follows. In Section 2, a brief review about the related work is presented. In Section 3, the proposed framework is described. The experiments conducted on three different datasets are presented in Section 4. The conclusions drawn and the future work are provided in Section 5.

2. Related work

In co-occurrence based document representation, there are three major steps. These steps are defining the features, selecting the best subset of these features and weighting the selected features. In this section, a literature review about the work carried out on these tasks is presented.

2.1. Definition of co-occurrence based features

The co-occurrence based features can be categorized into three groups, namely syntactic phrases, statistical phrases and termsets.

2.1.1. Syntactic phrases

Syntactic phrases are sequences of words ordered according to grammatical relations. Noun phrases, verb phrases and adjective phrases are typical syntactic phrases. The use of syntactic phrases for text classification was firstly studied by Lewis (1992b). He studied the use of BOW and syntactical phrases-based features separately and reported that syntactic phrases do not provide better scores compared to the BOW-based representation. Dumais et al. (1998) have observed that using syntactic phrases in addition to BOW generally degrades the performance achieved by using BOW alone. Scott and Matwin (1999) also noted that syntactic phrases do not provide a better representation compared to BOW. However, it is shown that voting over the outputs of the classifiers making use of BOW and phrase-based representation can provide better scores than the individual systems. This verifies that the phrases and BOW-based representations may complement each

other. The findings of Nastase et al. (2006) supported his idea. In particular, they studied the use of syntactically related pairs of words together with BOW and have shown that their approach provides improved accuracies compared to the BOW-based representation. More recently, Özgür and Güngör (2010) have shown that augmenting BOW with 37 lexical dependencies based features leads to significant improvements when compared to the BOW-based representation.

Although the use of grammatical relations between words is common to all of these studies, the types of the relations and the pruning levels considered to eliminate less frequent features are different. It can be argued that selecting a good subset of syntactic phrases is crucial for achieving improved performance scores by augmenting the BOW-based representation.

2.1.2. Statistical phrases

Statistical phrases, also known as n-grams, have been more extensively studied for text categorization. In this approach, sequences of n adjacent terms are used to define co-occurrence based features. The sequences of pairs (i.e. *bigrams*) and triples of words (i.e. *trigrams*) are generally considered where higher lengths are not found to be useful. Mladenic and Grobelnik (1998) have shown that the BOW-based representation can be successfully enriched by employing n-grams, $n \leq 3$. Similarly, Fürnkranz (1998) reported that sequences longer than three are not useful. Although the number of bigrams employed by Tan et al. (2002) to augment the BOW-based representation is 2% of the number of unigrams, improved classification performances are obtained. Instead of augmenting the BOW-based representation, Caropreso et al. (2001) kept the number of features used fixed where the bigrams are used to substitute some of the unigrams. However, they could not achieve promising results. Bekkerman and Allan (2004) studied the use of discriminative bigrams together with BOW. In their study, a bigram is a considered to be a candidate to be selected if its mutual information score is higher than the scores of the individual terms. They achieved improved scores compared to the BOW-based representation. Boulis and Ostendorf (2005) also studied the use of bigrams together with BOW on three datasets. They considered the additional information that each bigram brings when compared to its unigrams for choosing a good set of bigrams and reported improvements compared to BOW.

The use of varying length statistical phrases (multi-words) is also addressed. Zhang et al. (2008) studied the construction of multi-word based n-grams that have varying lengths. The multi-words are computed by comparing different sentences to find consecutive matching word sequences. However, the performance scores achieved were inferior to BOW. The similar problem is also addressed by Peng et al. (2013) where a context graph based approach is proposed to identify significant statistical phrases of arbitrary lengths. On the contrary, they reported significantly improved precision and recall scores compared to BOW, bigram and trigram based representations on two different datasets.

The common problem that is generally addressed in the use of statistical phrases is the selection of a good subset. Otherwise, a large set of additional features would be considered together with a large set of words which may lead to the problem of curse of dimensionality. The main difference among the existing studies is the criteria considered for selection. It can be concluded that the selection criteria are decisive regarding the performance of the categorization system. A review of the selection schemes widely utilized is presented in Section 2.2.

2.1.3. Termsets

In the termset-based approach, co-occurrences of different terms which are not necessarily adjacent is considered in defining novel features. In this approach, the terms do not need to form a syntactically meaningful sequence since their order is not important. In general, a subset of available terms is considered in defining termsets since all possible combinations of terms correspond to a huge set. For instance, Zaïane and Antonie (2002) employed pairs of frequent terms to define 2-termsets. By combining frequent terms and frequent 2-termsets, candidate 3-termsets are then generated. Association rules are computed to construct the resultant text classification system. Their simulation studies have shown that the results obtained are generally worse compared to the BOW-based representation. The study is later extended to employ the frequencies of the termsets during generating classification rules (Rak et al., 2005). Experimental results have shown that it is beneficial to use frequencies of termsets in text classification. Tesar et al. (2006) studied the use of both bigrams and 2-termsets. Based on their experiments, they argued that bigrams are more appropriate for text categorization. However, they reported that the use of termsets or bigrams do not provide any improvement to the BOW-based representation. Recently, Figueiredo et al. (2011) performed extensive experiments on the use of termsets for text categorization. In their study, individually discriminative terms are considered for defining termsets. A subset of the termsets obtained is then selected by applying a threshold on the document frequencies. The final set of 2-termsets to augment BOW is computed by selecting discriminative ones. A dominance score that is inversely proportional with the number of distinct classes the termset under concern appears is used for this purpose. They reported significantly better scores compared to BOW and bigrams-based representations.

The selection of termsets is even more crucial than n-grams. The main reason is that a termset is assumed to exist regardless of the order of the terms. Statistical and syntactical phrases are made up of adjacent terms which increase the probability of obtaining discriminative pairs. However, termsets may include terms which appear in different parts of the documents. We believe that these should be the major reasons for its being less attractive compared to the statistical and syntactical phrases-based approaches.

2.2. Selecting and weighting co-occurrence based features

The review presented in Section 2.1 clearly indicates that the selection of the co-occurring terms is an important problem for the text categorization task and various strategies are developed for this purpose. On the other hand, although the terms having higher discriminative power are ensured to contribute more to categorization by employing collection frequency factors in BOW-based systems, this is generally underestimated when co-occurrence based features are utilized. More specifically, either binary or term frequency based representation is generally employed when both terms and co-occurrence based features are used. In this section, we review in more detail the selection and weighting schemes that are available in the literature. Table 1 presents ten well-known/recent studies and the criteria used for selecting the co-occurring terms. We can categorize the criteria into two groups. The first group includes the supervised metrics MI, Kullback–Leibler (KL) divergence, IG, odds-ratio (OR) and χ^2 where the class labels of the documents are utilized. These are well known metrics for the general feature selection task (Ogura et al., 2011). *Dominance* that is defined as the conditional probability of a class given that the termset occurred also belongs to this group. The second group includes unsupervised measures which do not take into account the labels of the documents. These are *support* and term frequency (*tf*). Support, which is also known as the document frequency, is defined as the number of documents where a termset or phrase occurs. Using a threshold on the term frequency corresponds to specifying the minimum number of times that a termset or phrase must occur in the training set. It can be seen in the table that support is the

Table 1
Criteria considered for selecting termsets or phrases.

Study	*MI*	*KL*	*IG*	*OR*	χ^2	*Dominance*	*Support*	*tf*
Bekkerman and Allan (2004)	✓							
Caropreso et al. (2001)			✓	✓	✓		✓	
Figueiredo et al. (2011)						✓	✓	
Fürnkranz (1998)							✓	✓
Mladenic and Grobelnik (1998)				✓			✓	
Rak et al. (2005)						✓	✓	
Tan et al. (2002)			✓				✓	✓
Zaïane and Antonie (2002)						✓	✓	
Zhang et al. (2008)			✓					
Boulis and Ostendorf (2005)		✓						

Table 2
The weighting schemes considered for document representation when termsets or phrases are utilized.

Study	*Binary*	*tf*	*tf* × *idf*
Bekkerman and Allan (2004)	✓		
Caropreso et al. (2001)			✓
Figueiredo et al. (2011)	✓		
Fürnkranz (1998)	✓		
Mladenic and Grobelnik (1998)		✓	
Rak et al. (2005)		✓	
Tan et al. (2002)	✓		
Zaïane and Antonie (2002)	✓		
Zhang et al. (2008)	✓		
Boulis and Ostendorf (2005)		✓	✓

most popular. It should also be noted that, in majority of the studies, two or more measures are employed.

Table 2 presents the term weighting schemes utilized in the studies mentioned above. It can be seen that the most popular representations are term frequency and binary. When termsets are considered, the number of times each term of the termset occurs may be different. For instance, the first may occur only once whereas the second occurs more than ten times. In such cases, a new definition for the frequency of the termset is necessary. As a matter of fact, binary representation is generally used for termsets.

It should be noted that both symmetric and asymmetric collection frequency factors are developed for the BOW-based representation. Asymmetric factors consider the terms that mainly occur in the positive class as more important than those in the negative class where symmetric ones consider the terms that mainly occur in the negative class as valuable as those in the positive class. For instance, the relevance frequency (*RF*) is an asymmetric scheme defined as (Lan et al., 2009)

$$RF(t_i) = \log_2\left(2+\frac{A}{\max\{1,C\}}\right), \tag{1}$$

where A and C denote the number of positive and negative documents which contain the term t_i respectively. The multi-class odds ratio (MOR) is a symmetric term weighting scheme defined as (Chen et al., 2009; Erenel et al., 2011)

$$MOR(t_i) = \log_2\left(2+\max\left\{\frac{AD}{BC},\frac{BC}{AD}\right\}\right), \tag{2}$$

where B and D denote the number of positive and negative documents which do not contain t_i. Several other supervised term weighting schemes for BOW-based representation exist in the literature (Erenel et al., 2011). The majority of these schemes such as χ^2, odds-ratio, gain-ratio and information gain were originally proposed for feature selection (Debole and Sebastiani, 2003; Lan et al., 2009; Altay and Erenel, 2010). Erenel et al. (2011) studied the weighting behaviors of five of these schemes by analyzing their contour lines. In that study, they also proposed a novel weighting approach that is based on the occurrence probabilities of terms in different classes and compared their scheme with the other weighting schemes.

It should be noted that, since the BOW-based features are concatenated with the co-occurrence based ones, the use of the best-fitting weights for both co-occurrence and BOW-based features is necessary to obtain more discriminative composite feature vectors. On the other hand, the use of supervised weighting schemes taking into account the occurrences of the terms in different classes is not well studied in the case of co-occurrence based features. This study incorporates extension of our previous efforts on computing better term weights by using supervised techniques to weighting termsets.

3. Proposed framework

The proposed approach for selecting and weighting termsets of unordered word pairs, $\{t_i, t_j\}$ for binary text classification is based on the co-occurrence statistics of the individual terms in positive and negative classes. In other words, rather than focusing only on whether they both occur or not, the proposed framework also takes into consideration the cases where one of the terms appears but not the other. The main motivation is to employ the joint statistics of the terms for selecting and weighting termsets. Consequently, discriminative information that may exist in the occurrence of one term but not the other is quantified and utilized in document representation.

Consider the example illustrated in Fig. 1 where the positive class corresponds to "law" and includes two documents, d_1 and d_2. The negative class denoted by "$\overline{\text{law}}$" contains three documents, d_3, d_4 and d_5. Assume that there are two terms where t_1 denotes the term "tennis" and t_2 denotes "court". It can be seen that the positive documents do not include t_1. The BOW-based representation is presented in the second row of the figure where the first and second elements of the document vectors correspond to t_1 and t_2 respectively. In this example, without any loss of generality, we assumed that the weights of t_1 and t_2 are a and b respectively in all documents. In text categorization, the inner product is the most-widely used similarity measure during classification. Using this measure, it can be seen that the similarity of d_1 and d_2, d_1 and d_3, and d_1 and d_5 is the same. In other words, BOW is not able to differentiate between some positive and negative documents. The last row presents the proposed representation where the third feature corresponds to "t_2 occurs but not t_1". It is assumed that the weight of this feature is c when it is nonzero. In this case, the similarity of d_1 and d_2 is greater than the similarity of d_1 and d_3, and the similarity of d_1 and d_5. Consequently, the positive documents are more similar to each other than to the negative ones.

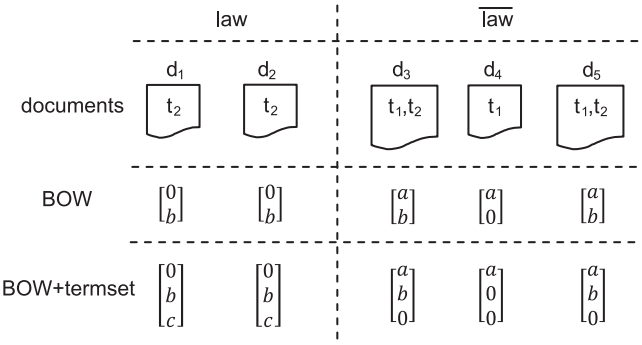

Fig. 1. An exemplar document classification problem illustrating the document vectors corresponding to BOW and an enriched representation (BOW+termset) including the feature "t_2 occurs but not t_1".

Table 3
The information elements employed in widely used selection and weighting schemes, A, B, C and D and their modified definitions, $\hat{A}$, $\hat{B}$, $\hat{C}$ and $\hat{D}$.

	Original definition		Modified definition
A:	The number of positive documents which include both t_i and t_j	$\hat{A}$:	The number of positive documents which include either or both of t_i and t_j
B:	The number of positive documents which do not include at least one of t_i and t_j	$\hat{B}$:	The number of positive documents which do not include any of t_i and t_j
C:	The number of negative documents which include both t_i and t_j	$\hat{C}$:	The number of negative documents which include either or both of t_i and t_j
D:	The number of negative documents which do not include at least one of t_i and t_j	$\hat{D}$:	The number of negative documents which do not include any of t_i and t_j

In order to implement such a representation, the information elements employed in widely used selection schemes, A, B, C and D, are firstly modified to take into account the occurrence of only one of the terms as presented in Table 3. It can be easily seen that the definition of occurrence is modified. More specifically, a termset is assigned a nonzero weight if *either* or *both* of the terms occur. For instance, $\hat{A}$ is the number of positive documents where at least one of the terms of the termset under concern appears. On the other hand, a termset does not occur if none of the terms appears in the given document. In the following context, the terms employed for defining a termset will be referred as *members* of the termset.

Consider the well-known selection scheme, χ^2 defined as

$$\chi^2 = \frac{N(AD-BC)^2}{(A+C)(B+D)(A+B)(C+D)}. \tag{3}$$

Replacing the original information elements with their modified forms, the χ^2 values of the termsets denoted by $\hat{\chi}^2$ can be computed as

$$\hat{\chi}^2 = \frac{N(\hat{A}\hat{D}-\hat{B}\hat{C})^2}{(\hat{A}+\hat{C})(\hat{B}+\hat{D})(\hat{A}+\hat{B})(\hat{C}+\hat{D})}. \tag{4}$$

It should be noted that the proposed information elements can be used with other selection schemes. In selecting a subset of termsets, as in almost all studies in the literature, the support values of the termsets are firstly computed. The termsets whose members do not co-occur in minimum of three different training documents are discarded. $\hat{\chi}^2$ is then used to sort the remaining termsets before selection.

After selecting an apriori specified number of termsets, their weights which are composed of two factors, namely the term frequency and the collection frequency factor are computed for all documents. The collection frequency factor of a termset depends on the member terms that appear in the document under concern. In other words, the collection frequency factor of a given termset is document dependent. With the use of this weighting scheme, individual and pairwise occurrences are separately evaluated in constructing the document vectors. Four new information elements defined for this purpose are presented in Table 4 where N^+ and N^- denote the total numbers of positive and negative training documents respectively, and $\overline{\{t_i,\overline{t_j}\}}$ denotes the complement of $\{t_i,\overline{t_j}\}$. These elements are used when only one of the members occurs as follows:

Table 4
The information elements employed in defining the weights corresponding to two different cases: t_i occurs but not t_j denoted by $\{t_i,\overline{t_j}\}$ (on the left) and t_j occurs but not t_i denoted by $\{\overline{t_i},t_j\}$ (on the right).

Term pair occurrence	Positive class	Negative class	Term pair occurrence	Positive class	Negative class
$\{t_i,\overline{t_j}\}$	P	Q	$\{\overline{t_i},t_j\}$	R	S
$\overline{\{t_i,\overline{t_j}\}}$	(N^+-P)	(N^--Q)	$\overline{\{\overline{t_i},t_j\}}$	(N^+-R)	(N^--S)

- P: The number of positive documents which include t_i but not t_j.
- Q: The number of negative documents which include t_i but not t_j.
- R: The number of positive documents which do not include t_i but include t_j.
- S: The number of negative documents which do not include t_i but include t_j .

For instance, if t_i occurs but not t_j, the information elements P, Q, (N^+-P) and (N^--Q) are considered in computing the termset weights. When both members occur, the information elements A, B, C and D are used. Consequently, the termset weights are defined by considering the appearing member term(s) and the corresponding information elements.

Consider the relevance frequency (RF) given in Eq. (1). The weight of the termset $\{t_i,t_j\}$ based on RF can be formulated as follows:

$$\widehat{RF}(\{t_i,t_j\}) = \begin{cases} \log_2\left(2+\dfrac{A}{\max\{C,1\}}\right) & \text{both } t_i \text{ and } t_j \text{ occur} \\ \log_2\left(2+\dfrac{P}{\max\{Q,1)\}}\right) & t_i \text{ occurs but not } t_j \\ \log_2\left(2+\dfrac{R}{\max\{S,1)\}}\right) & t_j \text{ occurs but not } t_i \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

Similarly, multi-class odds ratio (MOR) can be defined for termsets as follows:

$$\widehat{MOR}(\{t_i,t_j\}) = \begin{cases} \log_2\left(2+\max\left\{\dfrac{AD}{BC},\dfrac{BC}{AD}\right\}\right) & \text{both } t_i \text{ and } t_j \text{ occur} \\ \log_2\left(2+\max\left\{\dfrac{P(N^--Q)}{(N^+-P)Q},\dfrac{(N^+-P)Q}{P(N^--Q)}\right\}\right) & t_i \text{ occurs but not } t_j \\ \log_2\left(2+\max\left\{\dfrac{R(N^--S)}{(N^+-R)S},\dfrac{(N^+-R)S}{R(N^--S)}\right\}\right) & t_j \text{ occurs but not } t_i \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

It can be easily seen that individual or joint occurrences of the member terms of a termset are weighted separately. Consider the termset {"*tennis*", "*court*"} mentioned before. In this case, with the help of proposed weighting, the occurrence of "tennis" but not "court" may produce a large weight while the occurrence of "court" but not "tennis" is assigned a small weight.

The term frequency factor is computed for each termset as the sum of the member frequencies. Let tf_i and tf_j denote the term frequencies of the members in the document under concern. Then, the term frequency factor is computed as (tf_i+tf_j). The overall weight is finally obtained as the product of the two factors. For instance, using $\widehat{RF}(\{t_i,t_j\})$ as the collection frequency factor, the weight of the termset $\{t_i,t_j\}$ is computed as

$$w(\{t_i,t_j\}) = (tf_i+tf_j) \times \widehat{RF}(\{t_i,t_j\}) \tag{7}$$

Similarly, other collection frequency factors such as $\widehat{MOR}(\{t_i,t_j\})$ can be employed simply by replacing $\widehat{RF}(\{t_i,t_j\})$.

The document vectors are constructed by concatenating BOW and termset-based representations. The product of term frequency and collection frequency factor is also utilized in BOW-based representation. For instance, using *RF* as the collection frequency factor, the weight of the term t_i is computed as

$$w(t_i) = tf_i \times RF(t_i) \tag{8}$$

In the simulation experiments presented in the following section, the collection frequency factor is set to be the same in both BOW and termset-based representations.

4. Experiments

In all simulations, the F_1 score is used as the performance measure that is defined as the harmonic mean of precision (P) and recall (R) as

$$F_1 = \frac{2\times P\times R}{P+R},$$
$$P = \frac{TP}{TP+FP},$$
$$R = \frac{TP}{TP+FN} \tag{9}$$

where *TP*, *FP* and *FN* denote true positives, false positives and false negatives respectively. Macro and micro F_1 scores are generally used to compute the average performances on a given dataset. Macro F_1 is the average of the F_1 scores computed for each category (Sebastiani, 2002). On the other hand, micro F_1 score is computed using the total *TP*, *FP* and *FN* values when all categories are considered. Both macro and micro F_1 scores are used in performance evaluation of the systems implemented.

4.1. Datasets

Three widely used datasets are employed for evaluating the proposed framework. These are the ModApte split of top ten classes of Reuters-21578, 20 Newsgroups and OHSUMED. Reuters-21578 ModApte Top10 has a skewed category distribution. There are a total of 9980 news stories. It is a subset of Reuters, including the largest ten categories in the corpus (Debole and Sebastiani, 2004). 20 Newsgroups is a larger corpus of 20,000 newsgroup documents that are more uniformly distributed among twenty different categories. It is freely available at "people.csail.mit.edu/jrennie/20Newsgroups/". OHSUMED corpus which includes 20,000 medical abstracts adopted by Joachims is considered (Joachims, 1998). There are totally 23 categories, each corresponding to a different cardiovascular disease. Half of the corpus is used for training. For all datasets, the positive class is defined as the category under concern and the negative class is defined as the union of all documents in the other categories.

4.2. Experimental setup

The documents are firstly preprocessed for stopwords removal using SMART stoplist (Buckley, 1985) and then stemming is applied using the Porter stemming algorithm (Porter, 1980). The document lengths are normalized afterwards using cosine normalization approach. Due to the high dimensionality of document vectors, it is experimentally verified by various researchers that support vector machines (SVM) provide superior performance compared to various others such as naive Bayes (Lan et al., 2009). In our simulations, SVM^{light} toolbox with linear kernel (Joachims, 1998, 1999) and k NN are utilized. Default cost-factor value ($C=1/\text{avg}(\bar{x}^T\bar{x})$) that is the inverse of the average of the inner product values of the training data is employed for SVM. On several datasets, it is observed that the F_1 scores generally plateau after 5000 features when SVM is used (Lan et al., 2009). As a matter of fact, the top 5000 features ranked by χ^2 are used in the BOW-based representation for SVM.

It is well-known that k NN achieves its best scores on smaller number of features compared to SVM (Lan et al., 2009). Moreover, the best-fitting number of features and the value of k are dataset dependent. The macro F_1 scores of the BOW-based approach are computed for 100, 200, 400, 500, 1000 and 2000 terms and $k\in\{5,10,15,20,25,30\}$ and the best parameter values are determined. The number of terms are computed as 200, 100 and 100 respectively for Reuters-21578, 20 Newsgroups and OHSUMED. The best values of k are computed as 30, 5 and 5 respectively.

All pairwise combinations of the selected terms are considered for constructing termsets. After discarding the termsets with support less than three, the remaining pairs are ranked using $\hat{\chi}^2$ given in Eq. (4). For SVM, the top $K\in\{1,5,10,25,50,100,150,200,250,500,1000,2000,4000,5000,10000\}$ termsets are then concatenated with the BOW-based representation. For k NN, the top $K\in\{1,5,10,25,50,100,150,200,250,500,1000,2000\}$ termsets are utilized for this purpose.

4.3. Simulations

Figs. 2–4 present the macro F_1 and micro F_1 scores achieved using *RF* as the collection frequency factor for the term weights and $\widehat{RF}$ for the termset weights on Reuters-21578, 20 Newsgroups and OHSUMED respectively where SVM is employed as the classification scheme. The terms selected using χ^2 are employed as BOW-based features and $\hat{\chi}^2$ is utilized for termset selection. The reference scores obtained using the baseline BOW-based representation are shown by the dashed lines. It can be seen in the figures that the termsets are able to contribute to the scores on all three datasets, even when a few of them are considered. Although the performance of the proposed framework is higher than that of BOW for large number of termsets such as twice the number of terms in the BOW-based representation (i.e., 10000), there are some dataset based differences. For instance, the macro F_1 curves approach a plateau when 250 termsets are employed on Reuters-21578 and 20 Newsgroups datasets whereas further improvements are achieved as the number of termsets increases on OHSUMED. This clearly shows that the number of discriminative termsets is dataset dependent.

For k NN, the macro F_1 and micro F_1 scores achieved using *RF* as the collection frequency factor for the term weights and $\widehat{RF}$ for the termset weights on Reuters-21578, 20 Newsgroups and OHSUMED are presented in Figs. 5–7. As in the case of SVM, the termsets contribute to the scores on all three datasets. However, the highest scores are achieved using smaller numbers of features compared to SVM. The performance drops that occur as the number of termsets increase are mainly due to the inability of k NN to handle

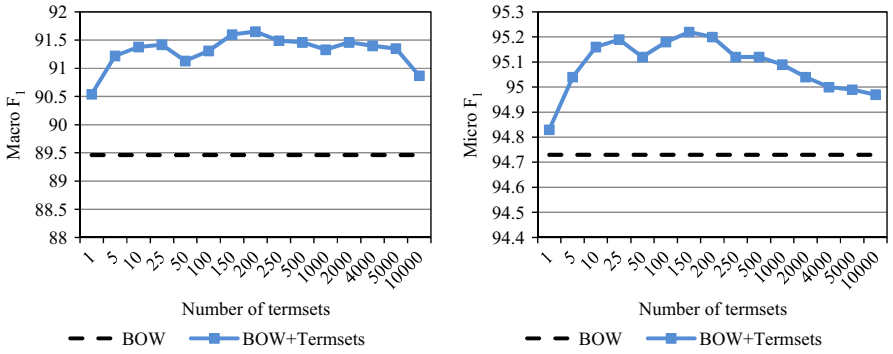

Fig. 2. The macro and micro F_1 scores achieved on Reuters-21578 by the proposed framework using *RF* and $\widehat{RF}$ as the collection frequency factors and SVM as the classification scheme.

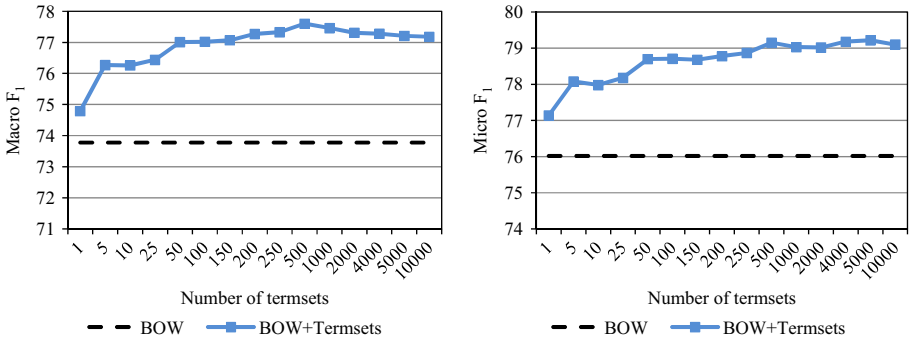

Fig. 3. The macro and micro F_1 scores achieved on 20 Newsgroups by the proposed framework using *RF* and $\widehat{RF}$ as the collection frequency factors and SVM as the classification scheme.

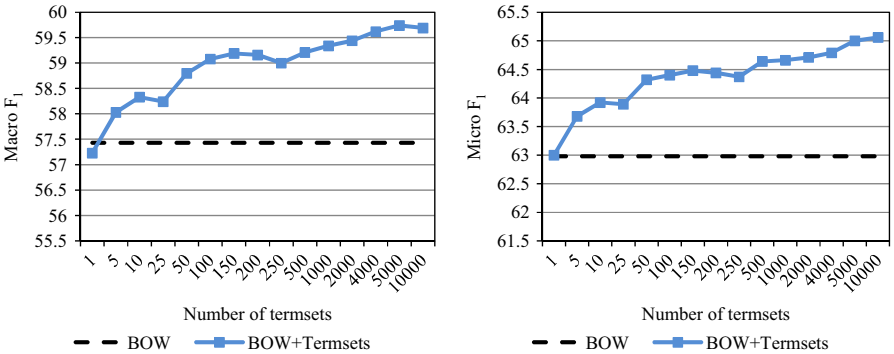

Fig. 4. The macro and micro F_1 scores achieved on OHSUMED by the proposed framework using *RF* and $\widehat{RF}$ as the collection frequency factors and SVM as the classification scheme.

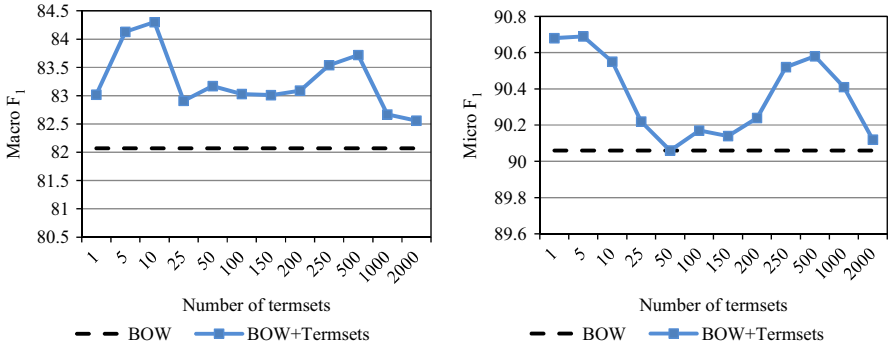

Fig. 5. The macro and micro F_1 scores achieved on Reuters-21578 by the proposed framework using *RF* and $\widehat{RF}$ as the collection frequency factors and *k* NN as the classification scheme.

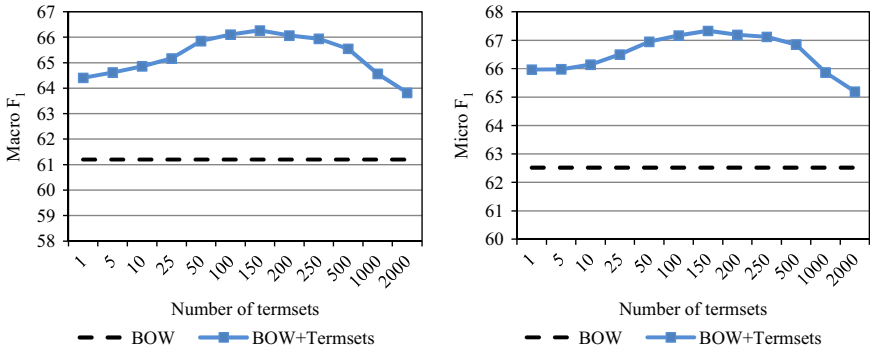

Fig. 6. The macro and micro F_1 scores achieved on 20 Newsgroups by the proposed framework using *RF* and $\widehat{RF}$ as the collection frequency factors and *k* NN as the classification scheme.

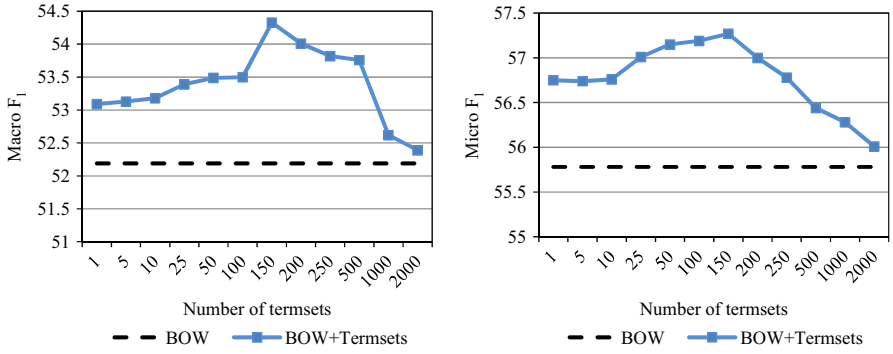

Fig. 7. The macro and micro F_1 scores achieved on OHSUMED by the proposed framework using *RF* and $\widehat{RF}$ as the collection frequency factors and *k* NN as the classification scheme.

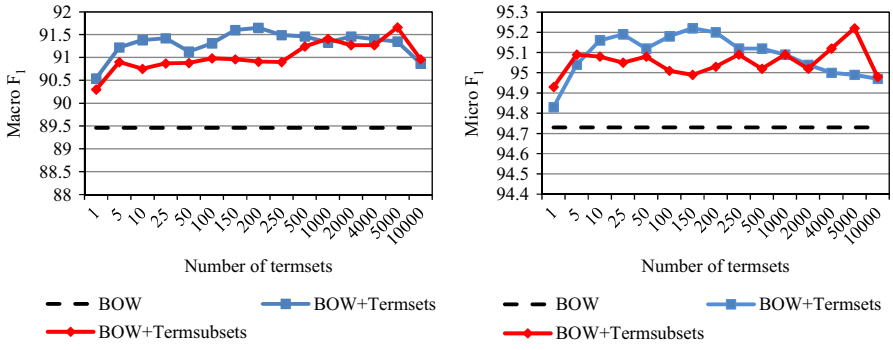

Fig. 8. The macro and micro F_1 scores achieved on Reuters-21578 by considering individual occurrences of terms but not their co-occurrence using *RF* and $\widehat{RF}_{ind}$ as the collection frequency factors.

large feature spaces (Lan et al., 2009). Comparing the performances of SVM and *k* NN, it can be seen that SVM provides superior scores than *k* NN in both BOW-based and proposed representations. Because of this, the experiments presented in the following context are conducted using SVM.

In order to verify the importance of using individual occurrence of one of the members but not the other, the termsets where both terms occur are discarded. Eq. (5) is modified for this purpose as follows:

$$\widehat{RF}_{ind}(\{t_i,t_j\}) = \begin{cases} 0 & \text{both } t_i \text{ and } t_j \text{ occur} \\ \log_2\left(2+\dfrac{P}{\max\{Q,1\}}\right) & t_i \text{ occurs but not } t_j \\ \log_2\left(2+\dfrac{R}{\max\{S,1\}}\right) & t_j \text{ occurs but not } t_i \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

Since the weight assigned to the co-occurrence of t_i and t_j is zero, we named these features as *termsubsets*. Figs. 8–10 present the macro F_1 and micro F_1 scores achieved using $\widehat{RF}_{ind}$ for the termsubset weights on Reuters-21578, 20 Newsgroups and OHSUMED respectively. The F_1 scores obtained using $\widehat{RF}\{t_i,t_j\}$ are also presented for comparison. The figures clearly demonstrate that the use of individual occurrences is fruitful on all three datasets. Considering the co-occurrences in addition to the individual occurrences provides further improvement on Reuters-21578 and OHSUMED. Because of this, in the following context, $\widehat{RF}$ will be considered for termset weighting.

The experiments are repeated by using *MOR* and $\widehat{MOR}$ as the collection frequency factors. Figs. 11 and 12 present the macro F_1 and micro F_1 scores achieved where the BOW-based representation employing *MOR* as the collection frequency factor is also presented as a reference. In the figures, it can be seen that improvements in F_1 scores are achieved as in the case of $\widehat{RF}$. On 20 Newsgroups dataset, the macro F_1 score decreases below the reference as the number of

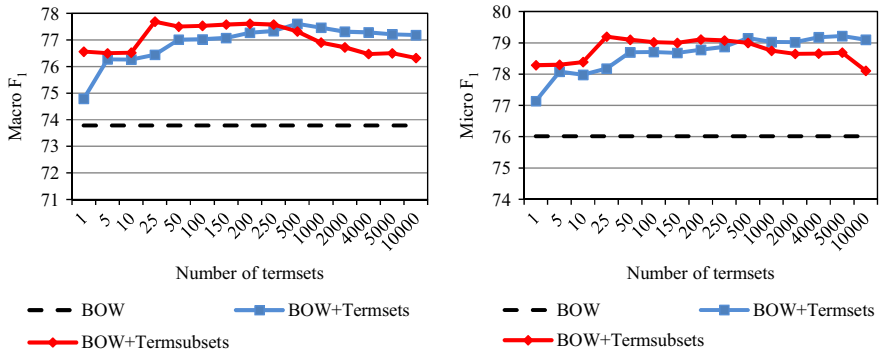

Fig. 9. The macro and micro F_1 scores achieved on 20 Newsgroups by considering individual occurrences of terms but not their co-occurrence using *RF* and $\widehat{RF}_{ind}$ as the collection frequency factors.

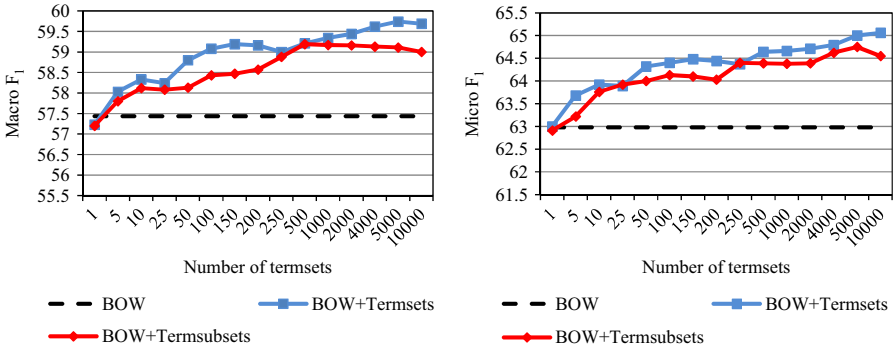

Fig. 10. The macro and micro F_1 scores achieved on OHSUMED by considering individual occurrences of terms but not their co-occurrence using *RF* and $\widehat{RF}_{ind}$ as the collection frequency factors.

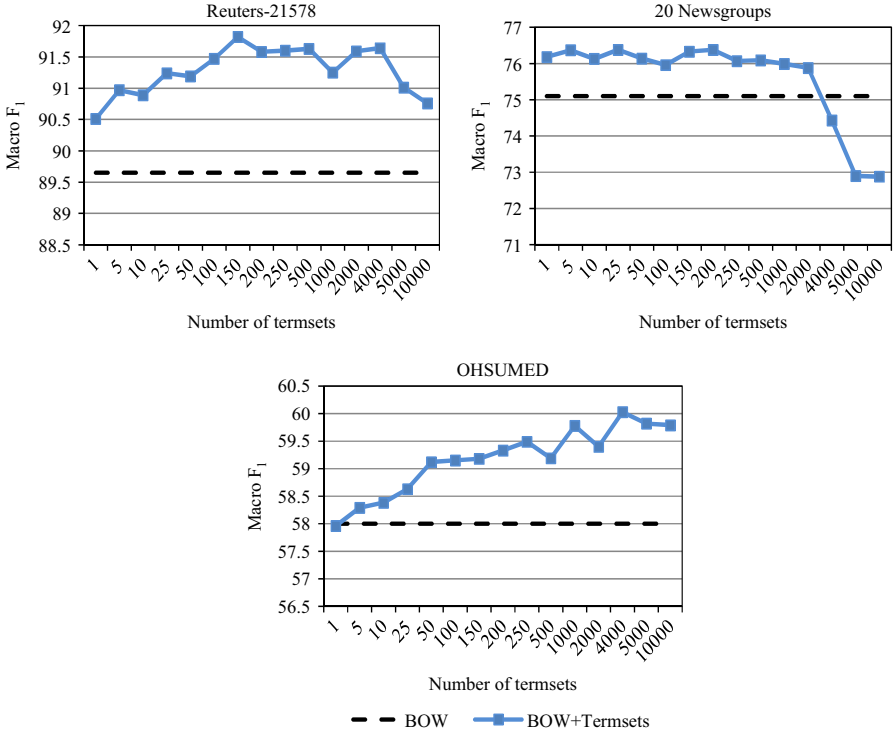

Fig. 11. The macro F_1 scores achieved by the proposed framework using *MOR* and $\widehat{MOR}$ as the collection frequency factors for BOW and termset-based representations respectively.

termsets increases to 4000. It should be noted that *MOR* is a symmetric scheme which considers the terms in the negative class as valuable as those in the positive. Hence, as more termsets are considered, it is likely that a large number of termsets which mainly appear in the negative class are employed. In order to verify this, the average values of $(\hat{A}/\hat{C})$ are computed for each dataset over all

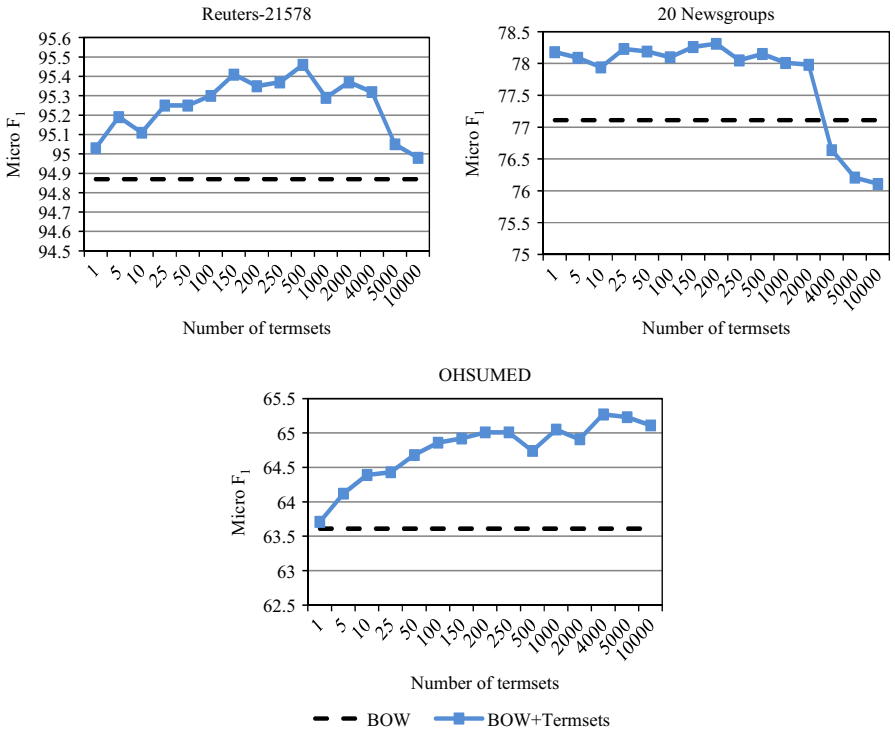

Fig. 12. The micro F_1 scores achieved by the proposed framework using *MOR* and $\widehat{MOR}$ as the collection frequency factors for BOW and termset-based representations respectively.

categories. It should be taken into consideration that the value of this expression decreases as more termsets are selected from the negative class. Table 5 presents the values obtained using the top ranked 1000 termsets and the termsets ranked between 9001 and 10000. It can be seen that the lower ranked termsets have lower values which means that they appear more frequently in the negative class compared to the higher ranked ones. For 20 Newsgroups dataset, $(\hat{A}/\hat{C}) < 1$ means that the termsets ranked between 9001 and 10000 appear in the negative class more frequently compared to the positive.

Remembering that the negative class includes documents from several categories that may not have common characteristics, it can be argued that the co-occurrence statistics of the member terms that mainly appear in the negative class may not always be reliable, leading to such a degradation. In fact, the degradation is mainly in the recall due to the increased number of false negatives. More specifically, when the use of 1000 and 10000 termsets together with BOW is compared, the macro recall is decreased from 67.47 to 64.32 due to the increase in the number of false negatives (from 119.30 to 130.55, on the average over all categories) where the macro precision remained almost unchanged. It can be concluded that the use of more negative features leads to the misclassification of increased number of positive documents. We also studied the use of 25,000 termsets for MOR. Both macro and micro F_1 scores slightly decrease for all three datasets when compared to 10,000 termsets. In particular, the macro and micro F_1 scores are obtained as 90.26 and 94.89 for Reuters-21578, 72.69 and 75.88 for 20 Newsgroups and, 59.66 and 64.98 for OHSUMED. However, the F_1 scores are still above the baseline in both Reuters-21578 and OHSUMED.

The experimental results presented above clearly demonstrate the effectiveness of the proposed framework. We conducted further experiments to investigate the relative performances of the selection schemes χ^2 and $\hat{\chi}^2$. Fig. 13 presents the macro F_1 scores achieved by utilizing these schemes for termset selection. *RF* and $\widehat{RF}$ are selected as the collection frequency factors for terms

Table 5
The average $(\hat{A}/\hat{C})$ values obtained using the top ranked 1000 termsets and ranked between 9001 and 10,000.

Dataset	Top 1000	Ranked between 9001 and 10000
Reuters-21578	7.43	3.91
20 Newsgroups	2.87	0.88
OHSUMED	3.00	1.51

in BOW and termsets respectively. As it can be seen from the figures, better scores are provided by $\hat{\chi}^2$ where the difference is less remarkable on Reuters-21578 dataset. In order to interrogate the comparable performance on this dataset, further experiments are performed. The $\hat{\chi}^2$ values of top 500 termsets selected by χ^2 and $\hat{\chi}^2$ are presented in Fig. 14. It can be seen in the figures that, on Reuters-21578, the termsets selected by χ^2 achieve higher $\hat{\chi}^2$ scores (around 1000) when compared to the other datasets. Because of this, they contribute to BOW-based representation on a similar order as those selected using $\hat{\chi}^2$. It can be concluded that, for the proposed document representation framework, $\hat{\chi}^2$ is ranking the termsets in a better way than χ^2.

The termsets selected using χ^2 and $\hat{\chi}^2$ are studied in terms of the number of times each word is employed in their construction. Fig. 15 presents the average number of times that the most frequently used ten terms appear as members when 5000 termsets are employed. It can be seen in the figure that a small set of terms are members in a large number of termsets when $\hat{\chi}^2$ is used. In other words, $\hat{\chi}^2$ emphasizes the co-occurrences of a small set of terms with the remaining ones. It can be seen in the figure that the terms ranked fifth or above are used a much fewer times, and hence a corresponding bar does not even appear. On the other hand, in χ^2, the most frequently used set of terms is larger. This means that $\hat{\chi}^2$ employs a wider set of different terms as members in the termsets.

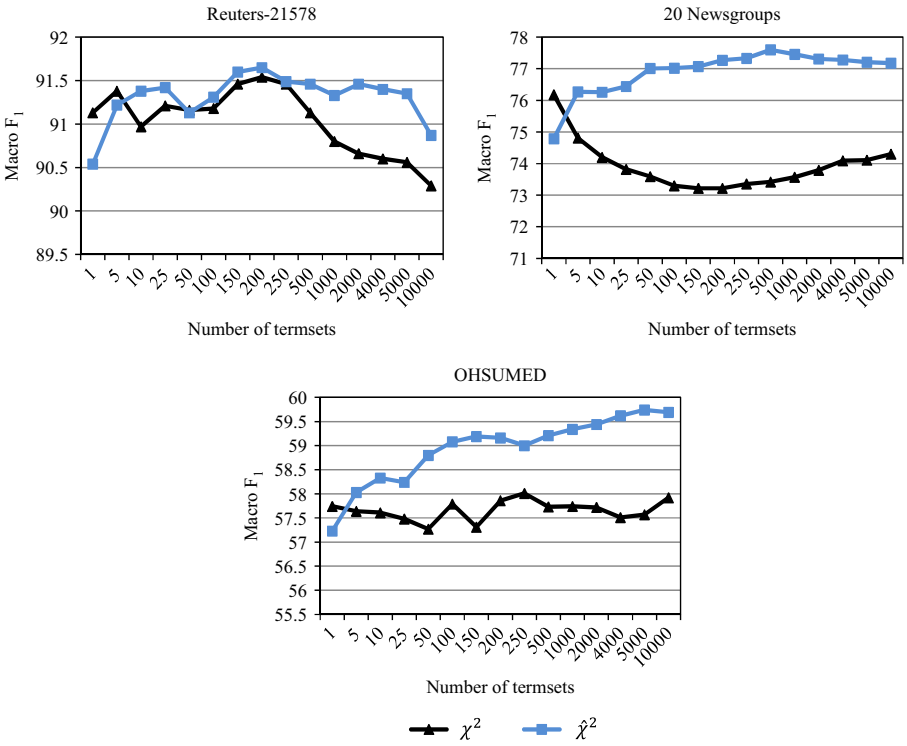

Fig. 13. The relative performances of the selection schemes χ^2 and $\hat{\chi}^2$ when *RF* and $\widehat{RF}$ are employed as the collection frequency factors for terms in BOW and termsets respectively.

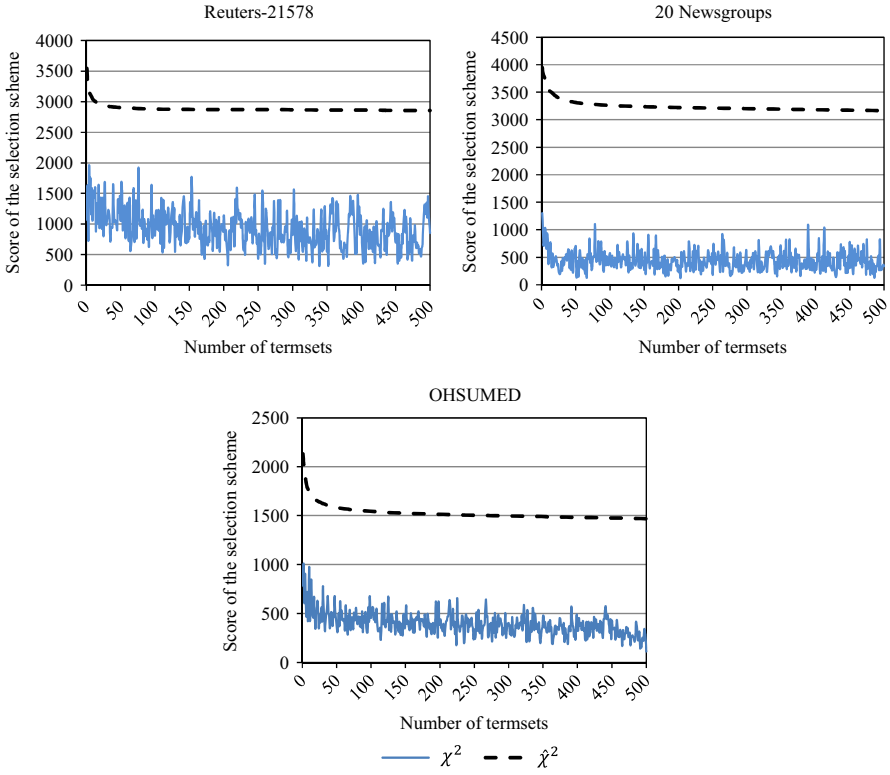

Fig. 14. The $\hat{\chi}^2$ values of top 500 termsets selected by χ^2 and $\hat{\chi}^2$.

The termsets selected using $\hat{\chi}^2$ are also investigated in terms of the total number of different terms utilized as a function of the number of termsets. Fig. 16 presents the average number of different terms used in the termsets selected over all categories using $\hat{\chi}^2$ as the termset selection scheme. On all three datasets, the average numbers of different terms employed increase almost linearly up to 500 termsets. The rate decreases as the number of termsets increases. For instance, on all datasets, approximately 500 different terms are employed in top ranked 500 termsets whereas, in the case of 5000 termsets, the number of different terms employed is approximately 3500 in 20 Newsgroups and OHSUMED.

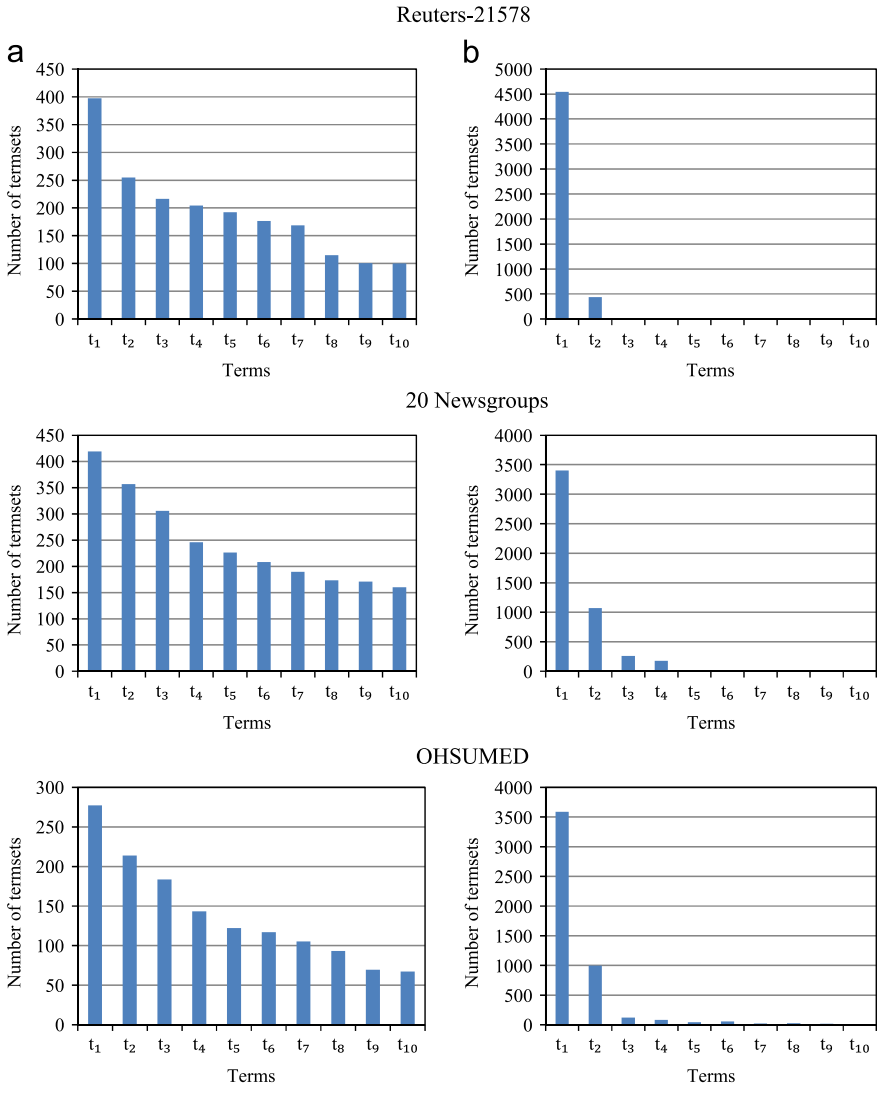

Fig. 15. The average number of times that the most frequently used ten terms appear as members when 5000 termsets are employed. (a) Using χ^2 for ranking and (a) Using $\hat{\chi}^2$ for ranking.

In our simulations, all 5000 terms utilized for BOW-based representation are considered for termset generation. This leads to $(5000 \times 4999)/2$ termsets which are more than twelve million possibilities. Although termset selection is done off-line during training, we studied the effect of using smaller number of terms for termset generation. More specifically, the use of 500, 1000, 2000, 3000 and 4000 terms that are top ranked using χ^2 is also studied for termset generation. It should be noted that, for 500 terms, the total number of different termsets are reduced to be $(500 \times 498)/2 = 124,750$ which is a much smaller number. Fig. 17 presents the macro F_1 scores achieved on three datasets. It can be seen that employing a large set of terms is beneficial where 4000 is the best-fitting number for all three datasets. We studied the training time required for termset selection when 5000 terms are utilized. On a 3.1GHz *i*5 processor, the total number of minutes needed for computing and ranking the termsets are computed as 38, 44 and 50 for the largest categories in Reuters-21578, 20 Newsgroups and OHSUMED respectively.

As stated in Section 2.2, the binary term weighting is generally considered when termsets are employed. We compared the performance of the proposed framework with the binary representation where the conventionally used scheme, χ^2 is utilized for termset selection. The results are presented in Fig. 18. The results for the proposed system using *RF* and $\widehat{RF}$ as the collection frequency factors and $\hat{\chi}^2$ for termset selection (denoted by BOW+Termsets(RF)) are also presented for comparison. It can be seen in the figure that, when binary representation is employed for term weighting, the use of termsets contributes to the BOW-based representation on two datasets, namely 20 Newsgroups and OHSUMED. However, the proposed scheme surpasses the binary representation based system for all different numbers of termsets considered on all datasets.

In order to asses the statistical significance of the improvements in the macro F_1 scores provided by the proposed approach, hypothesis tests are performed using the *t*-test approach. The null hypothesis is defined as "H0: mean of the improvement is equal to zero" and the alternative hypothesis is defined as "H1: mean of the improvement is greater than zero". The tests are performed for RF based weighting scheme using 500 termsets and BOW-based baseline system. The null hypothesis is rejected at significance levels of 0.05, with *p*-values 0.0400, 0.0035, 2.88×10^{-6} respectively for Reuters-21578, 20 Newsgroups and OHSUMED datasets.

The entire Reuters collection consist of 115 categories where Reuters-21578 is the subset of ten most frequent ones. In order to investigate the performance of the proposed scheme on less frequent classes, the experiments are repeated for all 115 categories. The experimental settings are the same as in the case of Reuters-21578. Fig. 19 presents the macro F_1 and micro F_1 scores

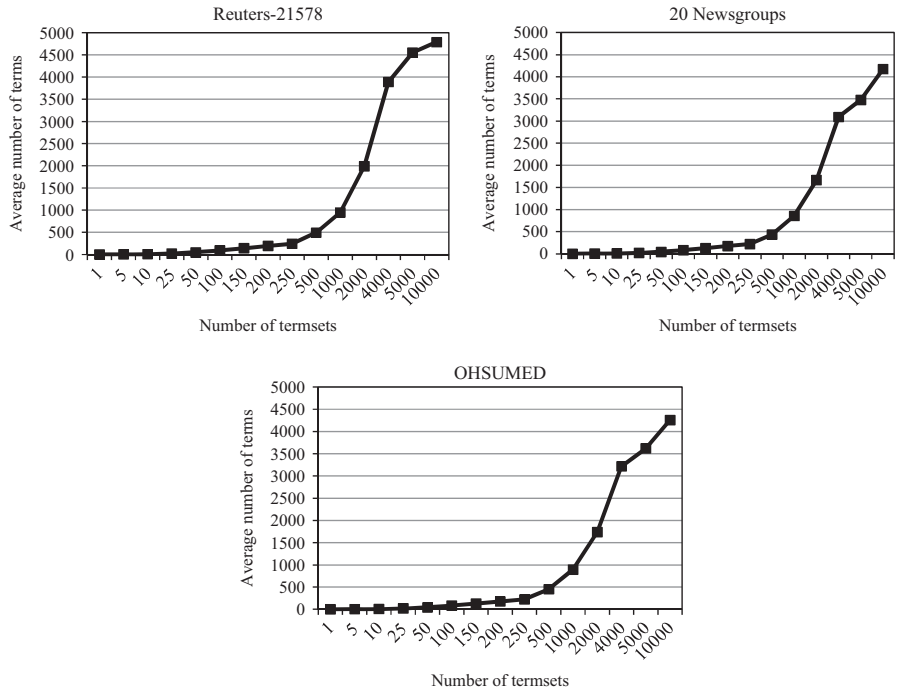

Fig. 16. The average number of different terms employed in the termsets selected using $\hat{\chi}^2$ as the termset selection scheme.

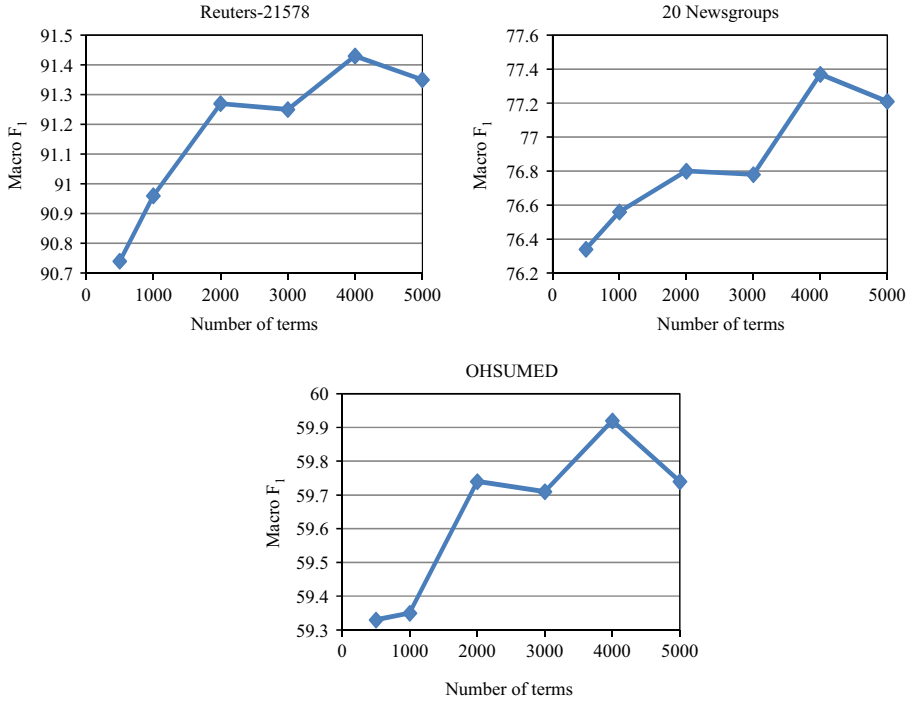

Fig. 17. The macro F_1 scores achieved on three datasets using different number of terms for termset generation using *RF* and $\widehat{RF}$ as collection frequency factors.

achieved using *RF* as the collection frequency factor for the term weights and $\widehat{RF}$ for the termset weights on the entire Reuters collection. Comparing Figs. 2 and 19, it can be seen that consistent improvements are achieved when less frequent categories are also considered. Fig. 20 presents the macro F_1 and micro F_1 scores achieved using $\widehat{RF}_{ind}$ for the termsubset weights on the entire Reuters collection. The F_1 scores corresponding to using $\widehat{RF}$ for termset weighting is also presented for comparison. The results clearly demonstrate that the use of individual occurrences is fruitful when less frequent categories are also considered.

The relative performances of the selection schemes χ^2 and $\hat{\chi}^2$ are also investigated on the entire Reuters collection. Fig. 21 presents the macro F_1 scores achieved by utilizing these schemes for termset selection. *RF* and $\widehat{RF}$ are selected as the collection frequency factors for terms in BOW and termsets respectively. As it can be seen from the figures, better scores are provided by $\hat{\chi}^2$. It should be noted that the difference between χ^2 and $\hat{\chi}^2$ is less remarkable on Reuters-21578 dataset when compared to 20 Newsgroups and OHSUMED as illustrated in Fig. 13. However, larger differences are observed when less frequent categories are also considered.

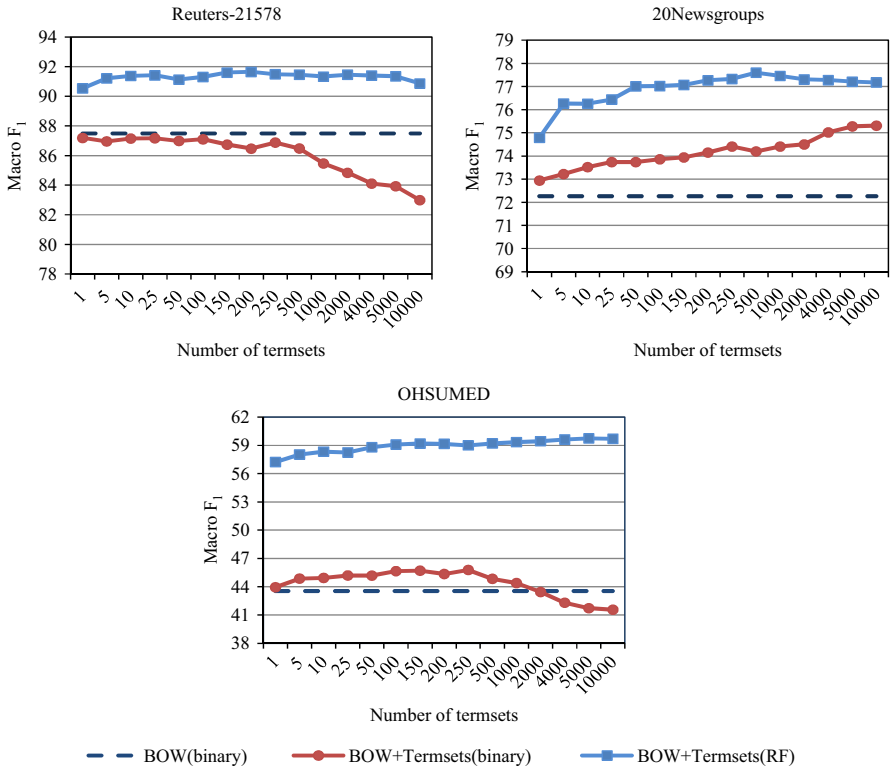

Fig. 18. The macro F_1 scores achieved on three datasets using χ^2 for both term and termset selection and binary term weighting. The performance of the proposed scheme is also presented for reference where *RF* is considered as the collection frequency factor.

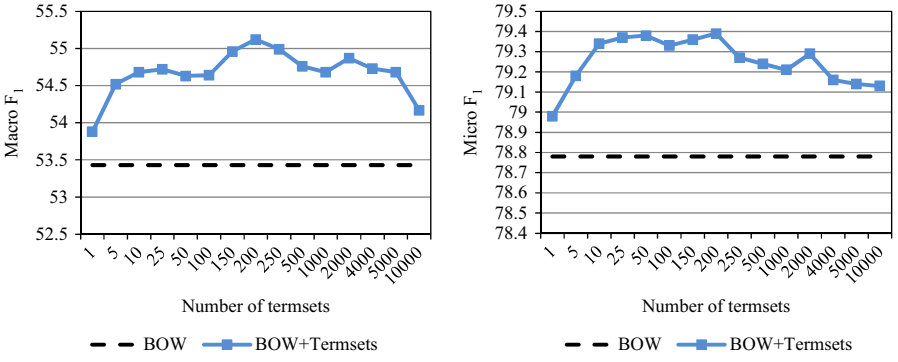

Fig. 19. The macro and micro F_1 scores achieved on the entire Reuters collection by the proposed framework using *RF* and $\widehat{RF}$ as the collection frequency factors and SVM as the classification scheme.

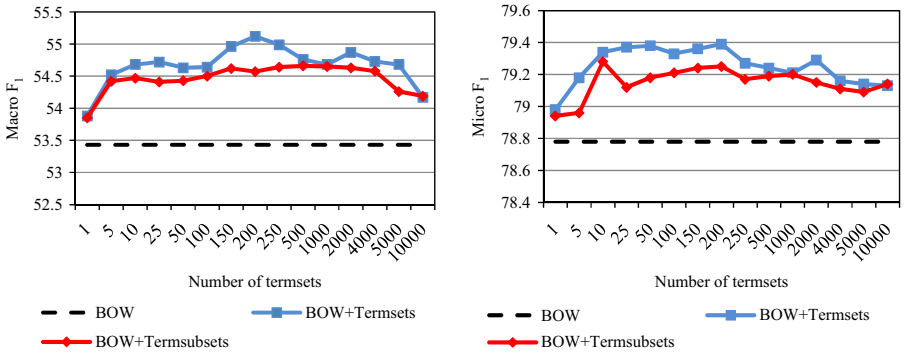

Fig. 20. The macro and micro F_1 scores achieved on the entire Reuters collection by considering individual occurrences of terms without their co-occurrences using *RF* and $\widehat{RF}_{ind}$ as the collection frequency factors.

We compared the performance of the proposed framework with the binary representation for the entire Reuters corpus. The results are presented in Fig. 22. The results for the proposed system using *RF* and $\widehat{RF}$ as the collection frequency factors and $\hat{\chi}^2$ for termset selection (denoted by BOW+Termsets(RF)) are also presented for comparison. It can be seen in the figure that, when

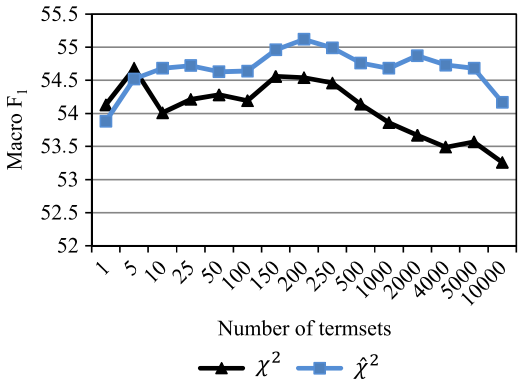

Fig. 21. The relative performances of the selection schemes χ^2 and $\hat{\chi}^2$ on the entire Reuters collection when *RF* and $\widehat{RF}$ are employed as the collection frequency factors for terms in BOW and termsets respectively.

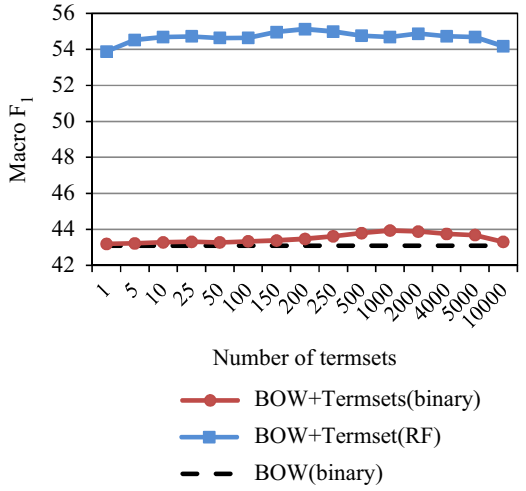

Fig. 22. The macro F_1 scores achieved on the entire Reuters collection using χ^2 for both term and termset selection and binary term weighting. The performance of the proposed scheme is also presented for reference where *RF* is considered as the collection frequency factor.

Table 6
The experimental setup, the best macro F_1 scores achieved and the corresponding numbers of termsets for SVM.

Experimental setup	Dataset	Number of termsets	Baseline (BOW)	Proposed scheme
SVM: $C=1/\text{avg}(\bar{x}^T\bar{x})$ SVM: kernel=linear Stopword list: SMART Stemming algorithm: Porter	Reuters-21578	200	89.46	91.65
Number of terms=5000 Term selection: χ^2 Termset selection: $\hat{\chi}^2$ Term weights: $tf_i \times RF(t_i)$ Termset weights: $(tf_i+tf_j)\times\widehat{RF}(\{t_i,t_j\})$	20 Newsgroups OHSUMED	500 5000	73.78 57.43	77.60 59.74

binary representation is employed for term weighting, the use of termsets has only slight contributions to the BOW-based representation. However, the proposed scheme provides remarkable improvements in the macro F_1 scores compared to the binary representation based system on the entire corpus.

5. Conclusions and future work

A novel framework is proposed to employ termsets having two member terms for binary text categorization. The joint occurrence statistics of the member terms are employed for termset selection and weighting. Eight new information elements are defined for this purpose. The experiments conducted on three widely used datasets have shown that, other than the co-occurrence of terms, the occurrence of only one of the members but not the other conveys discriminative information that helps to improve the performance of the BOW-based representation. It is also emphasized that the proposed approach can benefit from existing selection and weighting schemes simply by considering the proposed information elements.

The best-fitting number of termsets is observed to be dataset dependent. On the other hand, a few hundred termsets is observed to provide remarkable improvements on all three datasets. In particular, when 500 termsets are employed, it is shown that statistically significant improvements are achieved on all three datasets. It is also shown that using more terms for termset generation does not necessarily lead to better scores. The best-fitting numbers of termsets and the corresponding macro F_1 scores are presented in Table 6 when SVM is used as the classifier. The first column shows the design parameters that are common to all datasets. The numbers of termsets providing the highest macro F_1 scores are presented in the third column. It can be seen that, with the use of the proposed scheme, more than 2% improvement can be achieved in all three datasets.

There are several points that need to be further explored. In particular, choosing the best-fitting number of terms to compute the termsets is an important problem. Similarly, tuning the number of termsets to be employed is rather critical. On the other hand, the selection of terms and termsets is generally done independently. However, as emphasized by some researchers mentioned in Section 1, correlations between terms and termsets may exist, leading to a deterioration in the system performance. Taking into account the computational power available nowadays, it can be argued that developing better schemes for choosing the best subset of features from a bag of all terms and termsets should be one of the next endeavors of the researchers in this field.

References

Altınçay, H., Erenel, Z., 2010. Analytical evaluation of term weighting schemes for text categorization. Pattern Recognit. Lett. 31, 1310–1323.

Baker, D.L., McCallum, A.K., 1998. Distributional clustering of words for text classification. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98). ACM, New York, NY, USA, pp. 96–103.

Bekkerman, R., Allan, J., 2004. Using Bigrams in Text Categorization. Technical Report IR-408, Center of Intelligent Information Retrieval, UMass Amherst.

Boulis, C., Ostendorf, M., 2005. Text classification by augmenting the bag-of-words representation with redundancy compensated bigrams. In: Proceedings of the International Workshop on Feature Selection in Data Mining, in Conjunction with SIAM SDM-05, pp. 9–16.

Buckley, C., 1985. Implementation of the Smart Information Retrieval System. Technical report, Cornell University, Ithaca, USA.

Caropreso, M.F., Matwin, S., Sebastiani, F., 2001. A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization. In: Text Databases and Document Management, IGI Publishing, Hershey, PA, USA, pp. 78–102.

Chen, J., Huang, H., Tian, S., Qu, Y., 2009. Feature selection for text classification with naive Bayes. Expert Syst. Appl. 36, 5432–5435.

Debole, F., Sebastiani, F., 2003. Supervised term weighting for automated text categorization. In: Proceedings of the 2003 ACM Symposium on Applied Computing (SAC'03). ACM, New York, NY, USA, pp. 784–788.

Debole, F., Sebastiani, F., 2004. An analysis of the relative hardness of Reuters-21578 subsets. J. Am. Soc. Inf. Sci. Technol. 56 (6), 584–596.

Dumais, S., Platt, J., Heckerman, D., Sahami, M., 1998. Inductive learning algorithms and representations for text categorization. In: Proceedings of the Seventh International Conference on Information and Knowledge Management. ACM, pp. 148–155.

Erenel, Z., Altınçay, H., Varoğlu, E., 2011. Explicit use of term occurrence probabilities for term weighting in text categorization. J. Inf. Sci. Eng. 27 (3), 819–834.
Figueiredo, F., Rocha, L., Couto, T., Salles, T., Gonçalves, M.A., Meira, W., 2011. Word co-occurrence features for text classification. Inf. Syst. 36 (5), 843–858.
Fürnkranz, J., 1998. A Study Using n-Gram Features for Text Categorization. Technical Report OEFAI-TR-98-30, Austrian Research Institute for Artificial Intelligence, Austria.
Joachims, T., 1998. Text categorization with support vector machines: learning with many relevant features. In: Proceedings of the 10th European Conference on Machine Learning (ECML '98), Springer-Verlag, London, UK, UK, pp. 137–142.
Joachims, T., 1999. Making large-scale SVM learning practical. In: Schölkoph, B., Burges, C.J.C., Smola, A.J. (Eds.), Advances in Kernel Methods – Support Vector Learning. MIT Press, Cambridge, MA, pp. 169–184.
Lan, M., Tan, C.L., Su, J., Lu, Y., 2009. Supervised and traditional term weighting methods for automatic text categorization. IEEE Trans. Pattern Anal. Mach. Intell. 31 (4), 721–735.
Lewis, D.D., 1992a. An evaluation of phrasal and clustered representations on a text categorization task. In: Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '92). ACM, New York, NY, USA, pp. 37–50.
Lewis, D.D., 1992b. Representation and Learning in Information Retrieval (Ph.D. thesis). Amherst, MA, USA, UMI Order No. GAX92-19460.
Liu, Y., Loh, H.T., Sun, A., 2009. Imbalanced text classification: a term weighting approach. Expert Syst. Appl. 36, 690–701.
Mladenic, D., Grobelnik, M., 1998. Word sequences as features in text-learning. In: Proceedings of the 17th Electrotechnical and Computer Science Conference (ERK98), pp. 145–148.
Nastase, V., Shirabad, J.S., Caropreso, M.F., 2006. Using dependency relations for text classification. In: Proceedings of the 19th Canadian Conference on Artificial Intelligence.
Ogura, H., Amano, H., Kondo, M., 2011. Comparison of metrics for feature selection in imbalanced text classification. Expert Syst. Appl. 38 (5), 4978–4989.
Özgür, L., Güngör, T., 2010. Text classification with the support of pruned dependency patterns. Pattern Recognit. Lett. 31 (12), 1598–1607.
Peng, X., Yi, Z., Wei, X.Y., Peng, D.Z., Sang, Y.S., 2013. Free-gram phrase identification for modeling chinese text. Inf. Process. Lett. 113 (4), 137–144.
Porter, M.F., 1980. An algorithm for suffix stripping. Program 14 (3), 130–137.
Rak, R., Stach, W., Zaïane, O.R., Antonie, M.L., 2005. Considering re-occurring features in associative classifiers. Adv. Knowl. Discov. Data Min., 65–72.
Scott, S., Matwin, S., 1999. Feature engineering for text classification. In: Proceedings of the 16th International Conference on Machine Learning (ICML-99), Morgan Kaufmann, Bled, Slovenia, pp. 379–388.
Sebastiani, F., 2002. Machine learning in automated text categorization. ACM Comput. Surv. 34 (1), 1–47.
Tan, C.M., Wang, Y.F., Lee, C.D., 2002. The use of bigrams to enhance text categorization. In: Information Processing Management, vol. 38, pp. 529–546.
Tesar, R., Poesio, M., Strnad, V., Jezek, K., 2006. Extending the single words-based document model: a comparison of bigrams and 2-itemsets. In: Proceedings of the 2006 ACM Symposium on Document Engineering. ACM, New York, NY, USA, pp. 138–146.
Yang, J., Liu, Y., Zhu, X., Liu, Z., Zhang, X., 2012. A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization. Inf. Process. Manag. 48 (4), 741–754.
Zaïane, O.R., Antonie, M.L., 2002. Classifying text documents by associating terms with text categories. In: Proceedings of the 13th Australasian Database Conference (ADC '02), vol. 5, Australian Computer Society Inc., Darlinghurst, Australia, pp. 215–222.
Zhang, W., Yoshida, T., Tang, X., 2008. Text classification based on multi-word with support vector machine. Knowl.-Based Syst. 21 (8), 879–886.