Termset Selection and Weighting in Binary Text Classification

Dima Badawi

Submitted to the
Institute of Graduate Studies and Research
in partial fulfillment of the requirements for the Degree of

Doctor of Philosophy
in
Computer Engineering

Eastern Mediterranean University
June 2015
Gazimağusa, North Cyprus

Approval of the Institute of Graduate Studies and Research

Prof. Dr. Serhan Çiftçioğlu
Acting Director

I certify that this thesis satisfies the requirements as a thesis for the degree of Doctor of Philosophy in Computer Engineering.

Prof. Dr. Işık Aybay
Chair, Department of Computer Engineering

We certify that we have read this thesis and that in our opinion it is fully adequate in scope and quality as a thesis for the degree of Doctor of Philosophy in Computer Engineering.

Prof. Dr. Hakan Altınçay
Supervisor

Examining Committee

1. Prof. Dr. A. Aydın Alatan
2. Prof. Dr. Hakan Altınçay
3. Prof. Dr. Tolga Çiloğlu
4. Assoc. Prof. Dr. Ekrem Varoğlu
5. Asst. Prof. Dr. Nazife Dimililer

ABSTRACT

In this dissertation, a new framework that is based on employing the joint occurrence statistics of terms is proposed for termset selection and weighting. Each termset is evaluated by taking into account the simultaneous and individual occurrences of the terms within the termset. Based on the idea that the occurrence of one term but not the others 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 given termset is computed as a function of the terms that occur in the document under concern. 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-word approach to construct the document vectors.

As an extension to the proposed scheme, the use of cardinality statistics of the termsets is also considered for termset weight computation. More specifically, the cardinality statistics of the termsets that quantifies the number of member terms that occur in the document under concern is used for termset weighting. When employing termsets of length greater than two, cardinality-based weighting is observed to provide further improvements.

Keywords: Co-occurrence features, Cardinality statistics, Termset selection, Termset weighting, Document representation, Binary text classification.

ÖZ

Bu tezde, kelimelerin birlikte mevcudiyet istatistiklerine dayalı bir kelimeküme seçme ve ağırlıklandırma çerçevesi geliştirilmiştir. Her kelimeküme, içerdiği kelimelerin birlikte ve bağımsız olarak mevcudiyetleri dikkate alınarak değerlendirilmiştir. Bir kelimekümedeki kelimelerin sadece birinin mevcudiyetinin de ayırt edici değerli bilgi taşıyabileceği fikrinden yola çıkarak, geleneksel olarak kullanılan kelime seçme yöntemleri kelimeküme seçme amacıyla kullanılmak üzere güncellenmiştir. Benzer şekilde, verilen bir kelimekümenin ağırlığı, ilgili dökümanda yer alan kelimelerin bir fonksiyonu olarak tanımlanmıştır. Önerilen ağırlık kestirim yöntemi, kelimelerin tek başlarına ve birlikte mevcudiyetlerini ayrı ayrı değerlendirip dökümana bağlı ağırlıkların belirlenmesine olanak tanımaktadır. Önerilen kelimeküme-tabanlı gösterim ile kelime-çantası gösterimi birleştirilerek döküman vektörleri tanımlanmıştır.

Önerilen yaklaşımın bir uzantısı olarak, kelimekümelerin ağırlıklarının hesaplanmasında eleman sayısı istatistiklerinin kullanımı üzerinde de çalışılmıştır. Daha belirgin bir ifadeyle, kelimekümeler içerisindeki mevcut kelimelerin toplam sayıları ile ilgili bilgi içeren kelime sayısı istatistikleri, kelimeküme ağırlıklandırılmasında kullanılmıştır. İki kelimeden daha uzun kelimekümeler kullanıldığında, eleman sayısı tabanlı ağırlıklandırmanın daha fazla iyileştirme sağladığı gözlenmiştir.

Anahtar kelimeler: Birlikte mevcudiyet öznitelikleri, Eleman sayısı istatistikleri, Kelimküme seçme, Kelimeküme ağırlıklandırma, Döküman gösterimi, İkili metin sınıflandırma

ACKNOWLEDGEMENT

I thank all who in one way or another contributed in the completion of this thesis. First, I give thanks to God for protection and ability to do work.

I am also deeply thankful to my supervisor Prof. Dr. Hakan Altınçay for his deep insights and dedication to guide and help me through this thesis research. Without his creative, valuable supervision, this work would have encountered a lot of difficulties.

I would also like to thank my committee members, Prof. Dr. Aydın Alatan, Prof. Dr. Tolga Çiloğlu, Assoc. Prof. Dr. Ekrem Varoğlu, and Asst. Prof. Dr. Nazife Dimililer for serving as my committee members even at hardship.

Furthermore, my great thanks to my loving parents and my husband, for their support and encouragement through all these years in the Ph.D. program.

I would also like to thank all of my friends who supported me, and incented me to strive towards my goal.

Last but not least I would also like to thank my sisters, and brothers. They were always supporting me and encouraging me with their best wishes.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS/SYMBOLS

BOW	Bag of words
TC	Text categorization
SVM	Support vector machine
kNN	k-Nearest neighbor
tf	Term frequency
idf	Inverse document frequency
RF	Relevance frequency
χ^2	Chi-square
MI	Mutual information
MOR	Multi-class odds ratio
DF	Document frequency
OR	Odds ratio
GR	Gain ratio
KL	Kullback-Leibler divergence
N^+	The total number of documents in the positive category
N^-	The total number of documents in the negative category
N	The total number of documents
t_i	The i^{th} term
d	The number of unique features (terms) in the collection
n	The number of terms in the termset
$\mathbf{x_i}$	The i^{th} input vector
y_i	Class label, $1 \leq i \leq N$
R^d	Feature vector space with dimension d

$\mathbf{w}$	Weight vector
α_i	Lagrange multiplier
$k(\mathbf{x}, \mathbf{y})$	Kernel function
P	Precision
R	Recall
BEP	Break-even point
TP	True positives
FP	False positive
FN	False negative

Chapter 1

INTRODUCTION

1.1 Motivation

Automatic text classification is one of the key tasks in various problems such as spam filtering where 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) [1]. 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 Chi square (χ^2), Gini index or Information gain (*IG*) [2][3][4]. Then, the document vectors are constructed using frequencies of the selected terms which denote the number of times the terms occur 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 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 factor where the latter is used to take into account the relative importance of different terms [5]. For instance,

the inverse document frequency (idf) that considers less frequent terms as more important is used to define the term weights as ($tf \times idf$).

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 (ngrams) that are also known as statistical phrases is studied [6][7]. With the use of syntactic phrases, grammatical relations are also taken into consideration. ngrams are generally defined as consecutive occurrences of pairs (bigrams) or triples of terms (trigrams). The main idea is to use adjacent co-occurrences of different terms as novel features [6][8][9][10]. 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 [10]. 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) defined as the co-occurrences of terms having arbitrary order and position is also studied [11][12]. In this approach, if all terms appear in the document under concern, 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. 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 [13][14].

As in the BOW-based approach, selection of a good subset of co-occurrence based features is important and various criteria are studied for this purpose. In his study on the use of syntactic phrases, Lewis [7] 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 [9][10][15][16]. For instance, *IG* [9] and Mutual Information (*MI*) [10] 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 [17]. This criterion is also used for selecting a good subset of bigrams. For instance, in [14], it is 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. The authors 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 [15]. In fact, a weakness stated by Lewis is avoided by eliminating the rare words and the term dependencies with low occurrences. Figueiredo [11] 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 [11].

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 [3][18]. 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 [11]. Phrases-based features are defined as either binary or real-valued where, in the case of real-valued features, only the frequencies are generally considered for their weighting.

In this dissertation, a novel framework is proposed for selecting and weighting termsets. 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. As alternative approach, the joint occurrence statistics of the terms are utilized for termset selection and weighting where a termset may be assigned a nonzero weight even if all member terms do not appear in the document under concern. In other words, 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 may not always be essential for a termset to represent valuable information. As a matter of fact, instead of focusing only on 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 (i.e. 2-termsets) is firstly 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 that employs BOW representation.

The proposed approach for termset selection is also compared with the conventional selection schemes. More specifically, 2-termset selection using χ^2 and its adapted form are compared where remarkable improvements are observed.

The proposed framework is then extended to employ both 2-termsets and 3-termsets to enrich the BOW-based representation. The experiments have shown that, when 3-termsets are used together with 2-termsets, better scores are achieved when compared to employing BOW and 2-termsets only. However, superior scores are achieved only when small number of 3-termsets (50 or less) is considered and the performance is observed to degrade when more 3-termsets are used. It can be argued that the statistical information about the co-occurrences may not be reliably estimated as the length of the termsets increase. When the number of terms increases from two to three, the information elements employed to quantify the co-occurrence statistics increases from four to twelve, leading to reliable estimation problems. As a solution to this problem, we focused on employing the cardinality statistics of termsets for term weighting. In this approach, the termsets are weighted by taking into account the number of occurring terms within the termset. It is experimentally shown that more robust representations can be achieved. The use of 4-termsets is also addressed. It is observed that 4-termsets can contribute to the representation, providing improved scores on two of the three benchmark datasets.

In order to evaluate the proposed weighting scheme, further experiments are conducted. For instance, weighting bigrams is addressed. In this case, the 2-termsets are restricted to adjacent pairs of terms. In the conventional representation, a bigram assigned a non-zero weight if the member terms appear in the form an adjacent sequence. If both occur but they are not adjacent or one occurs but not the other, the bigram is said not to occur and its weight is zero. In these experiments, we considered assigning non-zero weights to bigrams even if only one of the terms occurs. The co-occurrence statistics of the terms that constitute bigrams is studied to develop a better weighting scheme. Experiments conducted on three widely used

benchmark datasets have shown that the proposed scheme contributes to the performance of BOW-based representation in two datasets and degrades for third. However, the scores are observed to be inferior on all three datasets when compared to the use of 2-termsets.

1.2 Thesis outline

The rest of this thesis is organized as follows. In Chapter 2, a detailed literature review about text categorization is presented. In particular, the text categorization is described and various techniques used to transform text documents into a vector form for automatic processing are studied. Several feature selection and weighting techniques and their importance in text categorization are addressed. It also provides a brief review about the most frequently used classifiers and datasets. Furthermore, it introduces document representation using termsets and bigrams.

Chapter 2 presents a brief literature review on text categorization including the efforts spent on employing co-occurrence based features. Chapter 3 presents the proposed framework for document representation using co-occurrence and cardinality statistics. Chapter 4 describes the experiments conducted for the evaluation of the proposed framework. Chapter 5 provides the conclusions drawn and discusses potential future research.

Chapter 2

LITERATURE REVIEW

The main aim in text classification (TC) is to compute the label of a given document as one or more from a predefined set of categories [19]. TC is a supervised learning problem, where labelled training data is used to compute a decision rule known as the classifier to predict the categories of unseen examples (test data). In general, document classification problems are formulated as binary where the positive class denotes the target category and the negative class includes all the remaining documents. After constructing a binary classifier for each category, they are combined to implement a multi-category classification system.

The main steps of text classification are preprocessing, document representation and classifier training. Figure 2.1 illustrates the main steps in preprocessing and representation of documents.

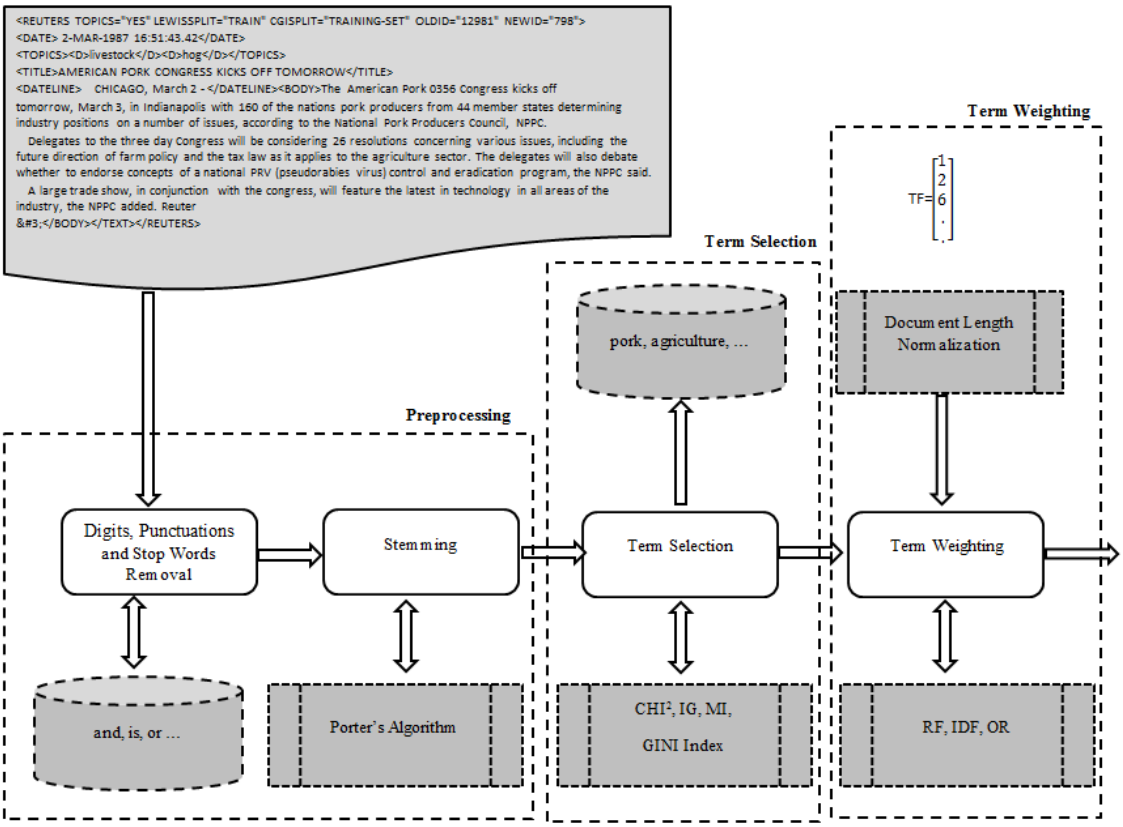


Figure 2.1: Main steps for preprocessing and representation of documents.

2.1 Preprocessing of documents

The initial step of preprocessing is the removal of the digits and punctuations. Any language includes words that have no semantic content by themselves. For instance, in English language, words like "the", "it" and "or" are not useful for text classification since they may occur in all categories in arbitrary frequencies. These so called stop words are generally removed [20]. SMART stop word list is the most widely used set of stop words to be discarded [21].

Another common preprocessing step is to perform stemming. It's a morphological normalization which has also been shown to improve results in information retrieval. Stemming uses simple rules of transformation to eliminate common inflexional

affixes. Porter stemmer is the most widely used stemming algorithm for English text [23].

2.2 Document representation

After the elimination of useless words and performing stemming, a set of candidate terms is left to be considered as features for constructing document vectors. The most common form of document representation in text categorization is the "bag-of-words" (BOW) where the features correspond to the words that appear at least once in the training corpus [24][25][26]. In general, the majority of these terms are not discriminative. As a matter of fact, a subset of them is employed in document representation. The selection of the best-fitting term set is a challenging problem and numerous measures are studied for this purpose. Experiments have shown that best performance scores are achieved when a few thousand terms are employed [18].

One of the widely explored approaches to enhance the BOW representation is the use of co-occurrence of words in addition to individual words. Although not remarkable in most cases, this approach improved categorization performance when compared to BOW. In co-occurrence based document representation, co-occurrences of the individually valuable terms are generally considered. After computing a good set of co-occurrence based features, the BOW-based vectors are enriched by these features. The co-occurrence based features can be categorized into three groups, namely syntactic phrases, statistical phrases and termsets.

2.2.1 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 [7]. 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. The authors in [27] have observed that using syntactic phrases in addition to BOW generally degrades the performance achieved by using BOW alone. Scott and Matwin [28] 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 phrases and BOW-based representation may complement each other. The findings in [29] 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, the authors in [15] 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.

2.2.2 Statistical phrases

Statistical phrases, also known as ngrams, have been more extensively studied for text categorization. In this approach, sequences of n adjacent terms (ngrams) 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 [6] have shown that the BOW-based

representation can be successfully enriched by employing ngrams, $n \leq 3$. Similarly, Fürnkranz [16] reported that sequences longer than three are not useful. Although the number of bigrams employed by Tan et al. [9] to augment the BOW-based representation is 2% of the number of the terms (unigrams), improved classification performances are obtained. Instead of augmenting the BOW-based representation, Caropreso et al. [8] 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. Authors in [10] 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. It should be noted that a detailed review of metrics used for co-occurrence based feature selection will be presented in Section 2.4. Boulis and Ostendorf [14] 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. [30] studied the construction of multi-word based ngrams 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 in [31] where a context graph based approach is proposed to identify significant statistical phrases of arbitrary lengths. On the contrary, they reported significantly improved performance 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.

2.2.3 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 [32] 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 [33]. Experimental results have shown that it is beneficial to use frequencies of termsets in text classification. Tesar et al. [12] 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. [11] 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 ngrams. 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.

2.3 Term selection

In practice, the number of terms retained after preprocessing and the numbers of phrases or termsets are on the order of thousands. In general, a subset of these features is employed for text classification. The reason for this is twofold. Firstly, some features may not convey discriminative information. Secondly, when all features are considered, the classifiers may overfit.

Feature selection aims to remove non-relevant features to reduce the dimensionality of the feature space and employ a discriminative set of features for classification. Various metrics are studied for this purpose. For instance, Yang and Pedersen [34] evaluated four different measures, namely document frequency thresholding (*DF)*,

information gain (*IG*), mutual information (*MI*) and Chi square statistic (χ^2) and Gini index (GI) [41]. These are well known feature selection metrics that are extensively studied in many domains.

Four information elements used in almost all selection schemes to quantify the importance of a given term, t are defined as follows:

A: The number of positive documents which include t.

B: The number of positive documents which do not include t.

C: The number of negative documents which include t.

D: The number of negative documents which do not include t.

The *DF* of a term is defined as the number of documents that include this term. In *DF* thresholding approach, the idea is that low frequency words are not helpful or relevant for class prediction. Using a predefined threshold, the terms whose document frequency is less than the threshold are removed [18].

Information gain measures the goodness of a term for class prediction by the evaluating the presence or absence of that term in different documents [34]. It is defined as [18]

$$IG = \frac{A}{N} log \frac{A \times N}{(A+C)(A+B)} + \frac{B}{N} log \frac{B \times N}{(B+D)(A+B)} + \frac{C}{N} log \frac{C \times N}{(A+C)(C+D)} + \frac{D}{N} log \frac{D \times N}{(B+D)(C+D)} \quad (2.1)$$

where *N* denotes the total number of training documents, i.e. $N = A+B+C+D$.

Mutual information is widely used in statistical language modeling [53]. It measures the dependence between two variables. In the field of text categorization, it is used to quantify the correlation among terms and categories. In other words, it measures the significance of a term for a particular category. It is defined as shown in Eq. 2.2 [35].

$$MI = log \frac{A \times N}{(A+C)(A+B)} \tag{2.2}$$

Chi square is used to measure the dependence of a term for both positive and negative classes. Strong association with the negative class also improves the χ^2 score [34]. Chi square value is determined as

$$\chi^2 = \frac{N(AD-BC)^2}{(A+C)(B+D)(A+B)(C+D)} \tag{2.3}$$

Gini Index is another feature selection metric which is an improved version of the one that is originally used to find the best split of attributes in decision trees [36]. It is defined as [37] [38]

$$Gini\ Index = \left(\left(\frac{A}{A+C}\right)^2 \left(\frac{A}{A+B}\right)^2 + \left(\frac{B}{B+D}\right)^2 \left(\frac{B}{A+B}\right)^2\right) * sign\ (AD - BC) \tag{2.4}$$

2.4 Co-occurrence based feature selection

The review presented in Section 2.2 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 under estimated 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 literature about co-occurrence based feature selection. Table 2.1 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*, *OR* and χ^2 where the class labels of the documents are utilized. 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 most popular. It should also be noted that, in majority of the studies, two or more measures are employed.

2.5 Term weighting

After term selection is done, the weights are calculated. The simplest weighting scheme is binary where the weight is one if the term occurs in a document and zero otherwise. As an alternative approach, tf can be used as the term weights. The term weights may also take into account the distribution of the terms in different classes. In order to realize this, the weights are defined as the products of two factors, namely the term frequency and the collection frequency factor.

Table 2.1: Criteria considered for selecting termsets or phrases.

Study	*MI*	*KL*	*IG*	*OR*	χ^2	Dominance	Support	tf
Bekkerman and Allan [10]	✓							
Caropreso et al. [8]			✓	✓	✓		✓	
Figueiredo et al. [11]						✓	✓	
Fürnkranz [16]							✓	✓
Mladenic and Grobelnik [6]				✓			✓	
Rak et al. [33]						✓	✓	
Tan et al. [9]			✓				✓	✓
Zaïane and Antonie [32]						✓	✓	
Zhang et al. [30]			✓					
Boulis andOstendorf [14]		✓						

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 inverse document frequency (idf) and the relevance frequency (*RF)* are asymmetric schemes. They are defined as [18]

$$idf = log\frac{N}{A+C} \tag{2.5}$$

$$RF = log\left(2 + \frac{A}{max(C,1)}\right) \tag{2.6}$$

where *A* and *C* denote the number of positive and negative documents which contain the term under concern, respectively. The multi-class odds ratio (*MOR*) is a symmetric term weighting scheme defined as [2][39]

$$MOR = log\left(2 + max(\frac{AD}{BC}, \frac{BC}{AD})\right) \tag{2.7}$$

where *B* and *D* denote the number of positive and negative documents which do not contain the corresponding term. Several other supervised term weighting schemes exists in the literature [39]. The majority of these schemes such as χ^2, odds ratio (*OR*), gain ratio and information gain were originally proposed for feature selection [5][18][40]. The authors in [39] 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 is recently verified that $tf \times RF$ achieves the best performance and outperforms other methods substantially on popular TC problems [18].

2.6 Weighting co-occurrence based features

Table 2.2 presents the weighting schemes utilized in the studies mentioned in Section 2.4. It can be seen that the most popular weighting schemes are term frequency and binary. When termsets are considered, the number of times each member term of the termset occurs may be different. For instance, the first may occur only once whereas the second occurs several 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.

Table 2.2: The weighting schemes considered for document representation when termsets or phrases are utilized.

Study	Binary	tf	$(tf \times idf)$
Bekkerman and Allan [10]	✓		
Caropreso et al. [8]			✓
Figueiredo et al. [11]	✓		
Fürnkranz [16]	✓		
Mladenic and Grobelnik [6]		✓	
Rak et al. [33]		✓	
Tan et al. [9]	✓		
Zaïane and Antonie [32]	✓		
Zhang et al. [30]	✓		
Boulis and Ostendorf [14]		✓	✓

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. However, 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.

2.7 Classification techniques

Support Vector Machines (SVM) [42], k-Nearest Neighbor (kNN) [43] and Naïve Bayes (NB) [44] are extensively studied for text categorization. Due to the high dimensionality of document vectors, it is experimentally verified by various researchers that SVM provides superior performance compared to various others including NB and kNN [18]. In [45], another comparison of performances of these classification techniques is presented. The results of this study also show NB provides inferior scores when compared to SVM and kNN.

As the reasons for the superiority of SVM in TC, Joachims pointed out the following arguments [19][46]:

- High-dimensional input space.
- Few irrelevant features: almost all features contain considerable information. He emphasized that a good classifier should combine many features and that aggressive feature selection may result in a loss of information.
- Sparse document vectors: despite the high dimensionality of the representation, each of the document vectors contain only a few non-zero elements.
- Linearly separability of most text categorization problems.

2.7.1 Support vector machines

SVM that is proposed originally proposed by Cortes and Vapnik [42] is a supervised learning approach based on the structured risk minimization principle. It is originally proposed for binary classification. SVM constructs the optimal hyperplane that separates the samples into two classes by maximizing the sum of its distances (margin) to the closest positive and negative vectors. As a result, the generalization error of the classifier is minimized.

Let $\{(\mathbf{x_i}, y_i)\}_{i=1}^{N}$ denotes the set of N training samples where $\mathbf{x_i} \in R^d$ is a real d-dimensional vector that belongs to the class $y_i \in \{-1, +1\}$.

Consider the linearly separable classes case where the separating hyperplane is desired to satisfy

$$\begin{cases} \mathbf{w^T x_i} + b \geq +1 & \text{if } y_i = +1 \\ \mathbf{w^T x_i} + b \leq -1 & \text{if } y_i = -1 \end{cases} \quad (2.8)$$

$\mathbf{w} = [w_1, w_2, \ldots, w_d]^{\boldsymbol{T}}$ is the weight vector for the hyperplane and b is the bias or offset from the origin.

The expressions above can be combined as

$$y_i(\mathbf{w}^T\mathbf{x_i} + b) \geq +1 \tag{2.9}$$

For linearly separable data, we find the separating hyperplane which maximizes the distance between it and the closest training sample. This distance or margin can be computed as $\frac{2}{\|\mathbf{w}\|}$. Maximizing $\frac{2}{\|\mathbf{w}\|}$ is equivalent to minimizing $\frac{\|\mathbf{w}\|^2}{2}$. Hence, the classification task that corresponds to computing the parameters $\mathbf{w}$ and b of the hyperplane can be formulated as the following constrained optimization problem:

$$min\frac{\|\mathbf{w}\|^2}{2}$$

Subject to: (2.10)

$$\begin{cases}\mathbf{w}^T\mathbf{x_i} + b \geq +1 & \text{if } y_i = +1\\ \mathbf{w}^T\mathbf{x_i} + b \leq -1 & \text{if } y_i = -1\end{cases}$$

The optimization problem can be re-written using Lagrange multipliers and the dual problem can be solved. Eq. 2.10 can be translated into the following form:

$$\begin{aligned} L_p &= \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{N} \alpha_i(y_i(\mathbf{w}^T\mathbf{x_i} + b) - 1) \\ &= \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{N} \alpha_i\, y_i(\mathbf{w}^T\mathbf{x_i} + b) + \sum_{i=1}^{N} \mathbf{x_i} \end{aligned} \tag{2.11}$$

where α_i is a Lagrange multiplier. The dual is to minimize L_p subject to the constraints that its gradients are set to zero as

$$\frac{\partial L_p}{\partial w} = 0 \rightarrow \mathbf{w} = \sum_{i=1}^{N} \alpha_i y_i \mathbf{x_i} \tag{2.12}$$

$$\frac{\partial L_p}{\partial b} = 0 \rightarrow \sum_{i=1}^{N} \alpha_i y_i = 0 \quad (2.13)$$

Lagrange multipliers are restricted to be non-negative *i.e.* $\alpha_i \geq 0$. The Lagrange multipliers are zero for all $\mathbf{x_i}$ except those lying on the hyperplanes which satisfy $y_i(\mathbf{w}^T\mathbf{x_i} + b) = 1$. There is one Lagrange multiplier for each training sample. The training samples for which the Lagrange multipliers are non-zero are called support vectors. Samples for which the corresponding Lagrange multiplier is zero can be removed from the training set without affecting the position of the final hyperplane.

By substituting the Eq. 2.12 and Eq. 2.13 in Eq. 2.11, the dual Lagrangian denoted by L_d is obtained as

$$L_d = \sum_{i}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j \, y_i y_j \mathbf{x_i}^T \mathbf{x_j} \quad (2.14)$$

After the Lagrange multipliers are computed, their values are used to find **w** and *b* and the class label of a test instance, **z** is calculated as

$$f(\mathbf{z}) = sign\ (\mathbf{w}^T\mathbf{z} + b) = sign\ (\sum_{i=1}^{N} \alpha_i y_i \mathbf{x_i}^T \mathbf{z} + b\ ) \quad (2.15)$$

In many cases, a separating hyperplane does not exist. If no hyperplane exists, it is possible to firstly map the sample points into a higher dimensional space using a non-linear mapping. That is, we choose a mapping φ: $R^d \rightarrow R^p$ where p is greater than d. We then seek a separating hyperplane in the higher dimensional space. This is equivalent to a non-linear separating surface in R^d.

When the mapping is taken into account, the expression $\mathbf{x_i}^T\mathbf{x_j}$ will be replaced by $\varphi(\mathbf{x_i})\varphi(\mathbf{x_j})$. If p is very large, this product could be difficult or expensive to

compute. However, this can also achieved using kernel function $\mathrm{k}(\mathbf{x_i}, \mathbf{x_j}) = \varphi(\mathbf{x_i})\varphi(\mathbf{x_j})$ [47], and we never need to know explicitly what φ is. Some examples of kernel functions are the polynomial kernel, $\mathrm{k}(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^{\boldsymbol{T}}\mathbf{y} + 1)^p$ and the radial basis function (RBF) kernel, $\mathrm{k}(\mathbf{x}, \mathbf{y}) = e^{||\mathbf{x}-\mathbf{y}||^2/2\sigma^2}$. In this case, the optimization problem becomes

$$\text{Max } \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i\alpha_j \, \mathrm{y_i y_j k}(\mathbf{x_i}, \mathbf{x_j}) \tag{2.16}$$

Subject to

$$\begin{cases} \alpha_i \geq 0 \\ \sum_{i=1}^{N} \alpha_i \mathrm{y_i} = 0 \end{cases} \tag{2.17}$$

After solving for **w** and *b*, we determine the class that the test vector **z** belongs using

$$f(\mathbf{z}) = \sum_i \alpha_i \mathrm{y_i k}(\mathbf{x_i}, \mathbf{z}) + b \, . \tag{2.18}$$

2.7.2 k-Nearest Neighbor classifier

The k-Nearest Neighbor rule (kNN), also called the majority voting k-nearest neighbor, is one of the oldest and simplest non-parametric techniques in the pattern classification literature. In this rule, the label of a test pattern is computed as the label of the majority of its k nearest neighbors in the training set [43].

The choice of k is essential in building the kNN model. In fact, k can be regarded as one of the most important factors of the model that can strongly influence the quality of predictions. Another important parameter is the distance measure.

Numerous measures of distance have been proposed. Cosine similarity is one of the most popular [48]. Cosine similarity measure computes the cosine of the angle between the vectors. Firstly each vector is normalized to a unit vector and then the inner product of the two vectors is calculated as

$$\mathrm{sim}(\boldsymbol{x},\boldsymbol{y}) = \cos(\theta) = \frac{\mathbf{x}^{T}\mathbf{y}}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|} \ . \qquad (2.19)$$

Another popular distance measure is Euclidean that is defined as [44]

$$\mathrm{d}(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{\mathrm{i}=1}^{\mathrm{d}}(\mathrm{a_i}(\mathbf{x}) - \mathrm{a_i}(\mathbf{y}))^2} \qquad (2.20)$$

where $\boldsymbol{x}$ and $\boldsymbol{y}$ are two instances in the d-dimensional space and $\mathrm{a_i}(.)$ is the value of the i^{th} attribute.

2.8 Performance measures

The most-widely used performance measures for text categorization are precision and recall [49]. Each level of recall is associated with a level of precision. In general, the higher the recall, the lower the precision, and vice versa. The point at which recall and precision are equal is called the break-even point (BEP), which is often used as a single summarizing measure for comparing results.

F measure [45] is another widely used evaluation measure that is defined as

$$\text{F measure} = \frac{1}{\tau\frac{1}{\mathrm{p}} + (1-\tau)\frac{1}{\mathrm{R}}} \ . \qquad (2.21)$$

τ determines the influence of precision and recall. In general, the value of τ is set to 0.5 to assign equal weights to both precision and recall. This particular value

corresponds to the frequently used F_1 measure defined as the harmonic mean of precision and recall which is defined as

$$F_1 = \frac{2 \times P \times R}{P+R} \tag{2.22}$$

where,

$$P = \frac{TP}{TP+FP} \tag{2.23}$$

$$R = \frac{TP}{TP+FN} . \tag{2.24}$$

TP, FP and FN denote true positives (correctly predicted positive documents), false positives (misclassified negative documents) and false negatives (misclassified positive documents) respectively.

In general, the F_1 score is reported as an average value. There are two ways for computing this average: macro average and micro average. For computing the macro F_1 score, the F_1 values of each category is determined and then averaged [1]. The total values of the true positive, true negative, false positive, and false negative scores obtained using all categories are considered are used to calculate the micro F_1 value. In text categorization, it is desirable to have higher F_1 scores by boosting both precision and recall.

2.9 Significance Tests

The performances of text categorization systems are empirically compared in general where precision; recall or F_1 scores achieved are the key parameters of the comparisons. In order to assess the statistical significance of the improvements in either of these scores provided by a novel scheme, hypothesis tests are commonly performed using the t-test approach [50][51]. In this test, the null hypothesis is defined as "*H*0 = mean of the improvement is equal to zero" and the alternative hypothesis is defined as "*H*1 = mean of the improvement is greater than zero". In order to perform the test, the value of the test statistic which follows normal distribution is firstly computed using the improvements achieved at the end of the experiments. If the resultant value falls in the rejection (or, critical) region, the null hypothesis is rejected. The critical region is specified by the level of significance, ρ which is typically selected as 0.05. We have adopted the t-test approach to evaluate the proposed approaches presented in this dissertation.

2.10 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.

2.10.1 Reuters Collection

The Reuters collection accounts for most of the experimental work done in text categorization so far. Thus, conducting experiments on this popular corpus provide meaningful comparison with the existing works. It consists of a set of news stories classified under categories related to economics. The entire Reuters collection consist of 115 categories. This dataset includes both large categories containing thousands of documents and small categories containing only a few.

Reuters-21578 ModApte Top10 split is the subset including ten most frequent categories. There are a total of 9,980 news stories [52]. This subset is more frequently used in text categorization research. Table 2.3 presents the categories and the numbers of training and test documents within each category.

Table 2.3: The number of training and test documents in each category of Reuters-21578.

Category	Number of training documents	Number of test documents
Earn	2877	1087
Acq	1650	719
Money-fx	538	179
Grain	433	149
Crude	389	189
Trade	369	117
Interest	347	131
Wheat	212	71
Ship	197	89
Corn	182	56

2.10.2 20 Newsgroups Collection

Another large benchmark data corpus is the 20 Newsgroups corpus. It is a collection of approximatly 20,000 newsgroup documents that are almost evenly divided among 20 discussion groups. Each document is labelled as one of the 20 categories corresponding to the name of the newsgroup that the document was posted to. It is freely available at "people.csail.mit.edu/jrennie/20Newsgroups/". Table 2.4 presents the categories and the numbers of training and test documents within each category.

Table 2.4: The number of training and test documents in each category of 20 Newsgroups

Category	Number of training documents	Number of test documents
alt.atheism	480	319
comp.graphics	584	389
comp.os.ms-windows.misc	572	394
comp.sys.ibm.pc.hardware	590	392
comp.sys.mac.hardware	578	385
comp.windows.x	593	392
misc.forsale	585	390
rec.autos	594	395
rec.motorcycles	598	398
rec.sport.baseball	597	397
rec.sport.hockey	600	399
sci.crypt	595	396
sci.electronics	591	393
sci.med	594	396
sci.space	593	394
soc.religion.christian	598	398
talk.politics.guns	545	364
talk.politics.mideast	564	376
talk.politics.misc	465	310
talk.religion.misc	377	251

2.10.3 OHSUMED Collection

The MEDLINE database is the largest component of PubMed (http://pubmed.gov). The OHSUMED collection is a subset MEDLINE. It includes records between the years 1987 and 1991. It contains 348,566 references out of a total of over 7 million, covering all references from 270 medical journals over a five-year period.

From the whole set of 50,216 abstracts in OHSUMED corpus, Joachims used the first 10,000 documents for training and the second 10,000 documents for testing [19]. This subset is more frequently utilized in text classification studies. There are totally 23 categories, each corresponding to a different cardiovascular disease. This data set is available at " http://disi.unitn.it/moschitti/corpora.htm". Table 2.5 presents the categories and the numbers of training and test documents within each category.

Table 2.5: The number of training and test documents in each category of OHSUMED

Category	Number of training documents	Number of test documents
Bacterial Infections and Mycoses	1000	1222
Virus Diseases	422	577
Parasitic Diseases	146	140
Neoplasms	2240	2780
Musculoskeletal Diseases	635	911
Digestive System Diseases	1247	1329
Stomatognathic Diseases	214	342
Respiratory Tract Diseases	1062	1397
Otorhinolaryngologic Diseases	275	291
Nervous System Diseases	1309	1904
Eye Diseases	348	410
Urologic and Male Genital Diseases	1026	1112
Female Genital Diseases and Pregnancy Complications	605	840
Cardiovascular Diseases	2222	2339
Hemic and Lymphatic Diseases	533	782
Neonatal Diseases and Abnormalities	415	496
Skin and Connective Tissue Diseases	649	755
Nutritional and Metabolic Diseases	739	816
Endocrine Diseases	458	438
Immunologic Diseases	1106	1456
Disorders of Environmental Origin	995	1345
Animal Diseases	226	219
Pathological Conditions, Signs and Symptoms	3997	4856

Chapter 3

DOCUMENT REPRESENTATION USING CO-OCCURRENCE STATISTICS OF THE MEMBER TERMS

The proposed framework is based on employing the joint occurrence statistics of terms for termset selection and weighting. Each termset is evaluated by taking into account the simultaneous or individual occurrences of the terms within the termset. More specifically, the selection and weighting of termsets is based on the co-occurrence statistics of the individual terms in the positive and negative classes. Rather than focusing only on whether all terms occur or not, the proposed framework also takes into consideration the cases where one of the terms appears but not the others. Consequently, discriminative information that may exist in the occurrence of one term but not the others is quantified and utilized in document representation. For a better understanding of the main idea, consider the example illustrated in Figure 3.1. Suppose that we have termsets of pairs (2-termsets, denoted by $\boldsymbol{t^2}$) 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 ω_1 and ω_2 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 ω_3 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.

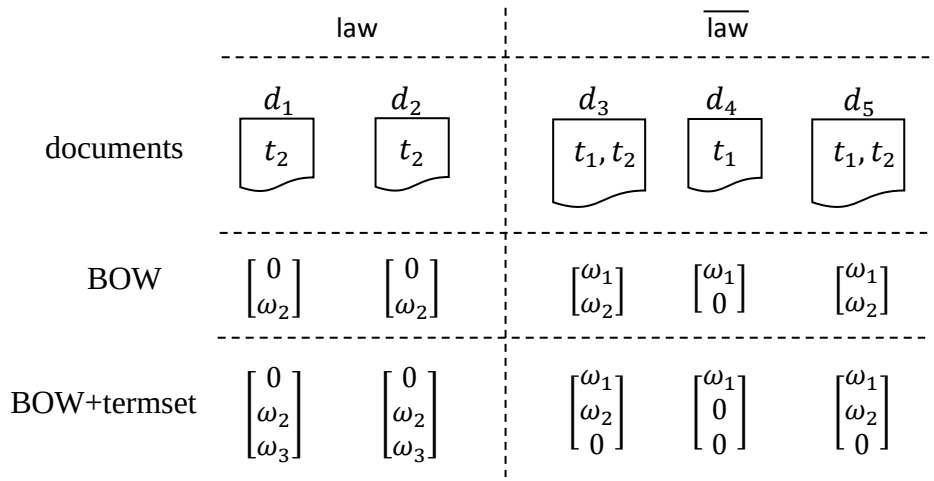


Figure 3.1: An exemplar document classification problem illustrating the document vectors corresponding to BOW and an enriched representation (BOW+termset) including the feature "t_2 occurs but not t_1".

In order to implement such a representation, the information elements employed in widely used selection schemes, A, B, C and D that are explained in Chapter 2 for the single terms are firstly modified to take into account the occurrence of only one of the terms in the termset as presented in Table 3.1. 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, $\widehat{A}$ 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.

Table 3.1: 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 all terms in the termset	$\hat{A}$:	The number of positive documents which include either one or more of the terms
B:	The number of positive documents which do not include at least one of the terms	$\hat{B}$:	The number of positive documents which do not include any of the terms
C:	The number of negative documents which include all terms in the termset	$\hat{C}$:	The number of negative documents which include one or more of the terms
D:	The number of negative documents which do not include at least one of the terms	$\hat{D}$:	The number of negative documents which do not include any of the terms

Consider the well-known selection scheme, χ^2 defined in Eq. 2.3. By 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

$$\hat{\chi}^2 = \frac{N(\hat{A}\hat{D}-\hat{B}\hat{C})^2}{(\hat{A}+\hat{C})(\hat{B}+\hat{D})(\hat{A}+\hat{B})(\hat{C}+\hat{D})}. \quad (3.1)$$

It should be noted that the proposed information elements can also be used with other selection schemes.

After selecting the termsets, their weights should be computed using the same philosophy. In particular, the weight of a termset is based on the occurrence statistics of the members. Consider the case of 2-termsets denoted by $\boldsymbol{t^2}$. Four new information elements are defined for this purpose which are presented in Table 3.2. $N^+$and N^-denote the total numbers of positive and negative training documents,

respectively. Let the event $\{t_i, \bar{t}_j\}$ denote the occurrence of t_i but not t_j and $\{t_i, \bar{t}_j\}$ denote the complement of $\{t_i, \bar{t}_j\}$. It can be seen in the table that *P* and *Q* denote the numbers of positive and negative documents which include t_i but not t_j, respectively. Similarly, *R* and *Q* denote the numbers of positive and negative documents which include t_j but not t_i, respectively.

Table 3.2: The information elements employed in defining the weights corresponding to two different cases: t_i occurs but not t_j denoted by $\{t_i, \bar{t}_j\}$ and t_j occurs but not t_i denoted by $\{\bar{t}_i, t_j\}$.

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

In computing the 2-termset weights, if t_i occurs but not t_j, the information elements P, Q, $(N^+ - P)$ and $(N^- - Q)$ are considered. 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. (2.5). The weight of the 2-termset, $\boldsymbol{t^2}$ denoted by $\widehat{RF}(\boldsymbol{t^2})$ is defined as

$$\widehat{RF}(\boldsymbol{t^2}) = \begin{cases} \log\left(2 + \frac{\mathrm{A}}{\max(\mathrm{C},1)}\right) & \{t_i, t_j\} \text{ occurs} \\ \log\left(2 + \frac{\mathrm{P}}{\max(\mathrm{Q},1)}\right) & \{t_i, \bar{t}_j\} \text{ occurs} \\ \log\left(2 + \frac{\mathrm{R}}{\max(\mathrm{S},1)}\right) & \{\bar{t}_i, t_j\} \text{ occurs} \\ 0 & \text{otherwise} \end{cases} \tag{3.2}$$

Similarly, the multi-class odds ratio (*MOR*) is defined for 2-termsets as

$$\widehat{MOR}(\boldsymbol{t^2}) = \begin{cases} \log\left(2 + \max\left(\frac{\mathrm{AD}}{\mathrm{BC}}, \frac{\mathrm{BC}}{\mathrm{AD}}\right)\right) & \{t_i, t_j\} \text{ occurs} \\ \log\left(2 + \max\left(\frac{\mathrm{P}(N^- - \mathrm{Q})}{(N^+ - \mathrm{P})\mathrm{Q}}, \frac{(N^+ - \mathrm{P})\mathrm{Q}}{\mathrm{P}(\mathrm{N}^- - \mathrm{Q})}\right)\right) & \{t_i, \bar{t}_j\} \text{ occurs} \\ \log\left(2 + \max\left(\frac{\mathrm{R}(N^- - \mathrm{S})}{(N^+ - \mathrm{R})\mathrm{S}}, \frac{(N^+ - \mathrm{R})\mathrm{S}}{\mathrm{R}(N^- - \mathrm{S})}\right)\right) & \{\bar{t}_i, t_j\} \text{ occurs} \\ 0 & \text{otherwise} \end{cases} \quad (3.3)$$

It can be easily seen that individual and joint occurrences of the member terms of a 2-termset are weighted separately. Consider the 2-termset including the members "tennis" and "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.

In order to verify the importance of using individual occurrence of only one of the members, discarding the 2-termsets where both terms occur is also studied. Eq. 3.2 is modified for this purpose as

$$\widehat{RF}_{ind}(\boldsymbol{t^2}) = \begin{cases} 0 & \{t_i, t_j\} \text{ occurs} \\ \log\left(2 + \frac{\mathrm{P}}{\max(\mathrm{Q},1)}\right) & \{t_i, \bar{t}_j\} \text{ occurs} \\ \log\left(2 + \frac{\mathrm{R}}{\max(\mathrm{S},1)}\right) & \{\bar{t}_i, t_j\} \text{ occurs} \\ 0 & \text{otherwise} \end{cases}. \quad (3.4)$$

Since the weight assigned to the co-occurrence of t_i and t_j is zero, we named these pairs as *termsubsets*.

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 $\left(tf_i + tf_j\right)$. The overall weight is finally obtained as the product of the two factors. For instance, using $\widehat{RF}(\boldsymbol{t^2})$ as the collection frequency factor, the weight of the termset $\boldsymbol{t^2}$ is computed as

$$w(\boldsymbol{t^2}) = \left(tf_i + tf_j\right) \times \widehat{RF}(\boldsymbol{t^2}) \tag{3.5}$$

Similarly, other collection frequency factors such as $\widehat{MOR}(\boldsymbol{t^2})$ and $\widehat{RF}_{ind}(\boldsymbol{t^2})$ can be employed simply by replacing $\widehat{RF}(\boldsymbol{t^2})$.

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

$$w(t_i) = tf_i \times RF(t_i) \tag{3.6}$$

The proposed framework can be easily extended to 3-termsets. The same set of information elements defined in Table 3.1 will be used. However, the co-occurrences of three terms will generate increased number of events. Hence, computation of weights is updated accordingly. Consider the information elements defined in Table 3.3. The weight of the 3-termset, $\boldsymbol{t^3}$ based on *RF* can be formulated as follows:

$$\widehat{RF}(\boldsymbol{t^3}) = \begin{cases} \log\left(2 + \frac{\mathrm{A}}{\max(\mathrm{C},1)}\right) & \{t_i, t_j, t_k\} \text{ occurs} \\ \log\left(2 + \frac{X_1}{\max(Y_1,1)}\right) & \{t_i, \bar{t}_j, \bar{t}_k\} \text{ occurs} \\ \log\left(2 + \frac{X_2}{\max(Y_2,1)}\right) & \{t_i, t_j, \bar{t}_k\} \text{ occurs} \\ \log\left(2 + \frac{X_3}{\max(Y_3,1)}\right) & \{t_i, \bar{t}_j, t_k\} \text{ occurs} \\ \log\left(2 + \frac{X_4}{\max(Y_4,1)}\right) & \{\bar{t}_i, t_j, t_k\} \text{ occurs} \\ \log\left(2 + \frac{X_5}{\max(Y_5,1)}\right) & \{\bar{t}_i, t_j, \bar{t}_k\} \text{ occurs} \\ \log\left(2 + \frac{X_6}{\max(Y_6,1)}\right) & \{\bar{t}_i, \bar{t}_j, t_k\} \text{ occurs} \\ 0 & \text{otherwise} \end{cases} \quad (3.7)$$

Then, the weight of the termset $\boldsymbol{t^3}$ is defined as

$$w(\boldsymbol{t^3}) = \left(tf_i + tf_j + tf_k\right) \times \widehat{RF}(\boldsymbol{t^3}). \quad (3.8)$$

The definition of longer termsets is possible. For instance, the 4-termset, $\boldsymbol{t^4}$ can be defined in a similar way by considering sixteen distinct events. The corresponding weights are obtained as $w(\boldsymbol{t^4}) = \left(tf_i + tf_j + tf_k + tf_l\right) \times \widehat{RF}(\boldsymbol{t^4})$. It should be noted that, as the length increases, the number of information elements to be computed increases exponentially. This may lead to unreliable estimates and hence poor representations. As a matter of fact, the choice of the maximum length is important.

Table 3.3: The information elements employed for co-occurrence based termset weighting.

Information element	Definition
X_1	The number of positive documents where $\{t_i, \bar{t}_j, \bar{t}_k\}$ occurs
Y_1	The number of negative documents where $\{t_i, \bar{t}_j, \bar{t}_k\}$ occurs
X_2	The number of positive documents where $\{t_i, t_j, \bar{t}_k\}$ occurs
Y_2	The number of negative documents where $\{t_i, t_j, \bar{t}_k\}$ occurs
X_3	The number of positive documents where $\{t_i, \bar{t}_j, t_k\}$ occurs
Y_3	The number of negative documents where $\{t_i, \bar{t}_j, t_k\}$ occurs
X_4	The number of positive documents where $\{\bar{t}_i, t_j, t_k\}$ occurs
Y_4	The number of negative documents where $\{\bar{t}_i, t_j, t_k\}$ occurs
X_5	The number of positive documents where $\{\bar{t}_i, t_j, \bar{t}_k\}$ occurs
Y_5	The number of negative documents where $\{\bar{t}_i, t_j, \bar{t}_k\}$ occurs
X_6	The number of positive documents where $\{\bar{t}_i, \bar{t}_j, t_k\}$ occurs
Y_6	The number of negative documents where $\{\bar{t}_i, \bar{t}_j, t_k\}$ occurs

In order to tackle with the potential problems of parameter estimation for longer termsets, the proposed framework is extended to use the *cardinality statistics* instead of the co-occurrences. More specifically, the cardinalities of the events that occur are taken into account to update the weighting of the termsets. It should be noted that the same set of information elements defined in Table 3.1 are used for termset selection. In particular, Eq. (3.1) is employed. Assume that the termset of length *n* is represented by $\boldsymbol{t^n}$. If we consider *RF* as the weighting scheme, the weight of $\boldsymbol{t^n}$ based on the cardinality statistics, $\widetilde{RF}(\boldsymbol{t^n})$ is computed as

$$\widetilde{RF}(\boldsymbol{t^n}) = \begin{cases} \log\left(2 + \frac{\mathrm{P_1}}{\max(\mathrm{Q_1},1)}\right) & 1 \text{ term from } \boldsymbol{t^n} \text{ occurs} \\ \log\left(2 + \frac{\mathrm{P_2}}{\max(\mathrm{Q_2},1)}\right) & 2 \text{ terms from } \boldsymbol{t^n} \text{ occur} \\ \vdots & \vdots \\ \log\left(2 + \frac{\mathrm{P_n}}{\max(\mathrm{Q_n},1)}\right) & \text{all } n \text{ terms in } \boldsymbol{t^n} \text{ occur} \\ 0 & \text{otherwise} \end{cases}, \quad (3.9)$$

where the information elements utilized are as defined in Table 3.4.

Table 3.4: The information elements employed for cardinality-based termset weighting.

Information element	Definition
P_1	The number of positive documents which include one term from $\boldsymbol{t^n}$
P_2	The number of positive documents which include two terms from $\boldsymbol{t^n}$
P_n	The number of positive documents which include n terms from $\boldsymbol{t^n}$
Q_1	The number of negative documents which include one term from $\boldsymbol{t^n}$
Q_2	The number of negative documents which include two terms from $\boldsymbol{t^n}$
Q_n	The number of negative documents which include n terms from $\boldsymbol{t^n}$

As in the case of co-occurrence statistics based weighting, the term frequency factor is computed as the sum of the member frequencies and the overall weight is finally obtained as the product of the two factors. For instance, using $\widetilde{RF}(\boldsymbol{t^n})$ as the collection frequency factor, the weight of the termset $\boldsymbol{t^n}$ is computed as

$$w(\boldsymbol{t^n}) = (tf_1 + tf_2 + \cdots + tf_n) \times \widetilde{RF}(\boldsymbol{t^n}) \quad (3.10)$$

It should be noted that the co-occurrence statistics of two or more terms are expected to contribute more to the representation if they are not independent. Since the member terms may be placed in arbitrary locations, it can be argued that many termsets will include independent members. Therefore, additional constraints can be

applied for termset selection. For instance, termsets may be replaced by *ngram*s, all of which form a subset of all termsets. In this case, the selection strategy should be updated while applying the same scheme for weighting. In this study, weighting of ngrams is also addressed. As an alternative to the conventional approach that takes into account the adjacent occurrences of the terms for weighting, we employ the joint occurrence statistics of the terms constituting the bigrams for this purpose. More specifically, based on the hypothesis that discriminative information may also exist in the occurrence of one term but not the other, the proposed scheme also employs the individual occurrence statistics of the terms for computing the weights of the corresponding ngrams. The document vectors are then constructed by concatenating the weight vectors of terms (unigrams) and ngrams.

Assume that $\boldsymbol{b^n}$ denotes an ngram of length *n*. For *n=2*, $\boldsymbol{b^2} =< t_i,t_j >$ is said to occur if both t_i and t_j appear in the document under concern in an adjacent form in the given order. The information elements used in the selection of ngrams are given in Table 3.5.

Table 3.5: The information elements employed for selection of ngrams.

Information element	Definition
A'	The number of positive documents which include $\boldsymbol{b^n}$
B'	The number of positive documents which do not include $\boldsymbol{b^n}$
C'	The number of negative documents which include $\boldsymbol{b^n}$
D'	The number of negative documents which do not include $\boldsymbol{b^n}$

Assume that *RF* is selected as the collection frequency factor. Consider the case of bigrams. Let P, Q, R, S be defined as given in Table 3.2. Let *X* denote the number of positive documents which include both t_i and t_j but do not include $\boldsymbol{b^2}$. In other words, X corresponds to the number of documents that include both terms but they never appear in consecutive form. Similarly, let *Y* denote the number of negative documents which include both t_i and t_j but do not include $\boldsymbol{b^2}$. Then $RF'(\boldsymbol{b^2})$ is defined as

$$RF'(\boldsymbol{b^2}) = \begin{cases} \log\left(2 + \frac{A'}{\max(C',1)}\right) & < t_i,t_j > \text{occurs} \\ \log\left(2 + \frac{\mathrm{X}}{\max(\mathrm{Y},1)}\right) & \{t_i, t_j\} \text{ occurs but } < t_i,t_j > \text{does not occur} \\ \log\left(2 + \frac{\mathrm{P}}{\max(\mathrm{Q},1)}\right) & \{t_i, \bar{t}_j\} \text{ occurs} \\ \log\left(2 + \frac{\mathrm{R}}{\max(\mathrm{S},1)}\right) & \{\bar{t}_i, t_j\} \text{ occurs} \\ 0 & \text{otherwise} \end{cases} \quad (3.10)$$

It should be noticed that the constraint of adjacent occurrence is applied during their selection. After the bigrams as selected, the partially occurred bigrams may also be assigned non-zero weights. The term frequency factor is computed for each bigram as the sum of the member frequencies as before. Assume that tf_i and tf_j denote the term frequencies of the members of a particular bigram in the document under concern. Then, the term frequency factor of the bigram is computed as (tf_i+tf_j). Hence, the weight of the bigram becomes $(tf_i + tf_j) \times RF'(\boldsymbol{b^2})$.

The selection and weighting of ngrams can be easily extended to *n >2*. However, it is omitted due to obtaining inferior results compared to 2-termsets as presented in the following chapter.

Chapter 4

EXPERIMENTS

In this chapter, the proposed selection and weighting schemes are evaluated on three widely used datasets. The experimental results are also compared with the state-of-the-art text categorization schemes.

4.1 Experimental setup

In our simulations, both SVM and kNN are considered. Before computing the term and termsets weights, digits and punctuation marks are deleted, the stop words are removed using SMART list and stemming is applied using Porter stemmer. Then, the document lengths are normalized using cosine normalization. The normalized forms of the term frequencies are used to compute the final forms of the weights of the terms and termsets. After the document vectors are computed, classifiers are trained using the training data.

In our simulations, SVM^{light} toolbox with linear kernel is used for training and evaluation of the SVM classifier [19][46]. The default cost-factor value ($C = {}^{1}/_{avg(\bar{x}^T\bar{x})}$) that is the inverse of the average of the inner product values of the training data is employed. On several datasets, it is observed that the F_1 scores generally plateau after 5000 features when SVM is used [18]. It is also shown that χ^2 provides the better scores for 5000 features when compared to the others. As a matter of fact, the top 5000 features ranked by χ^2 are used in the BOW-based representation for SVM.

In general, kNN achieves its best scores on smaller number of features compared to SVM [18]. 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 numbers 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. We used cosine similarity measure for kNN in all our experiments.

All combinations of the selected terms are considered for constructing termsets. After discarding the termsets with support less than three, the remaining termsets are ranked and weighted according to proposed weighting framework. The first set of experiments are done for the 2-termsets, and then extended for 3-termsets and 4-termsets to be employed together with the 2-termsets.

For SVM, the top $v \in \{1,5,10,25,50,100,150,200,250,500,1000,2000,4000,5000, 10000\}$ termsets are concatenated with the BOW-based representation. For kNN, the top $v \in \{1,5,10,25,50,100,150,200,250,500,1000,2000\}$ termsets are utilized for this purpose.

4.2 BOW-based classification

The macro and micro F_1 scores obtained for the baseline BOW-based representation are presented in Table 4.1 for both SVM and kNN. The relative differences can be explained by the differences in the datasets characteristics. For instance some categories may contain longer or shorter documents on the average, and there may be domain specific differences.

Table 4.1: The macro and micro F_1 scores obtained for the baseline BOW-based representation.

Dataset	SVM		kNN	
	macro F_1	micro F_1	macro F_1	micro F_1
Reuters-21578	89.46	94.73	82.07	90.06
20 Newsgroups	73.78	76.02	61.2	62.52
OHSUMED	57.43	62.98	52.19	55.78

Figures 4.1, 4.2 and 4.3 present the F_1 scores achieved using BOW representation for each category of Reuters-21578, 20 Newsgroups and OHSUMED, respectively. The average number of terms in each category is also presented. Notice that the vertical axis is common to both scores. It can be seen that a general correspondence does not exist between average document lengths and F_1 scores. As a matter of fact, after normalizing the lengths using cosine normalization, document length differences are not taken into consideration in the experiments conducted on the use of termsets.

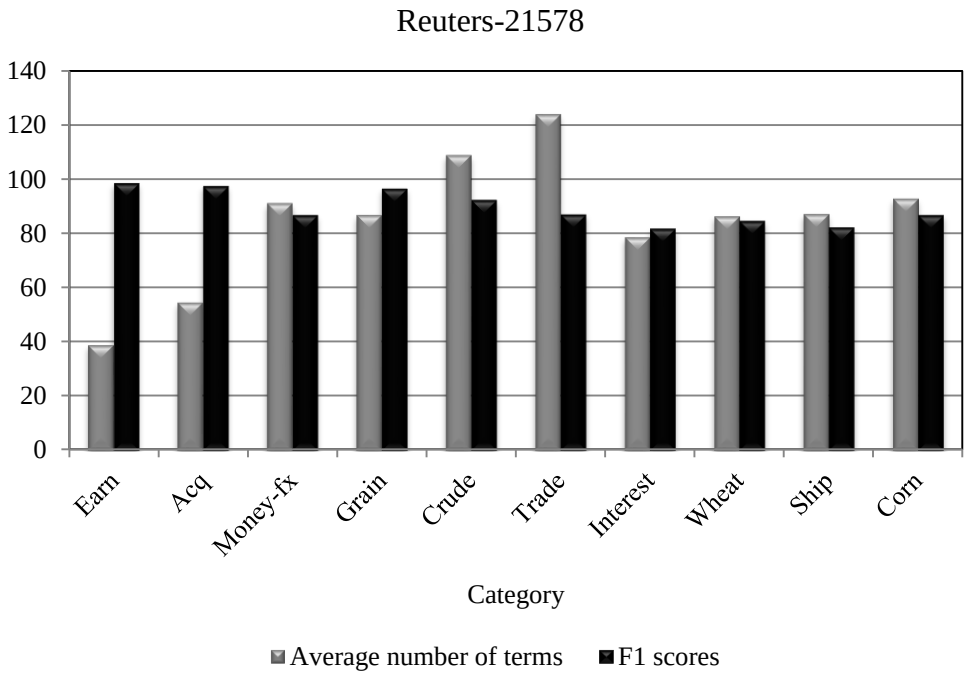


Figure 4.1: The F_1 scores achieved using BOW representation and the average number of terms for each category of Reuters-21578.

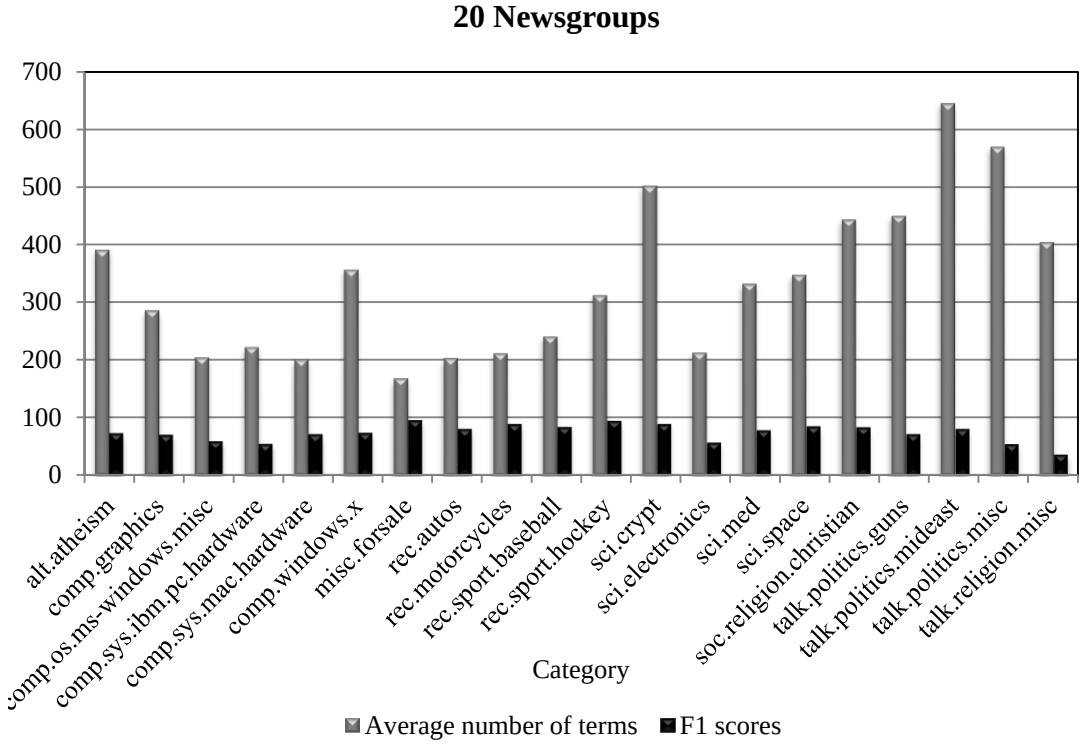


Figure 4.2: The F_1 scores achieved using BOW representation and the average number of terms for each category of 20 Newsgroups.

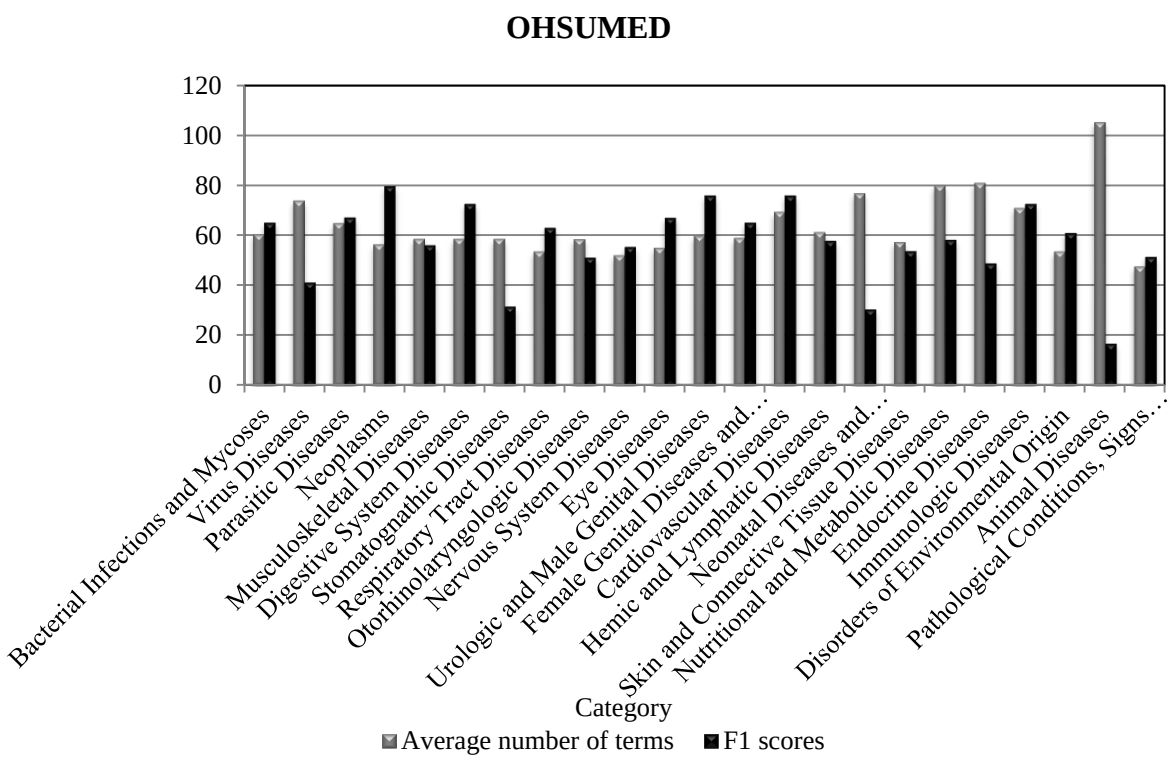


Figure 4.3: The F_1 scores achieved using BOW representation and the average number of terms for each category of OHSUMED.

4.3 2-Termset selection and weighting using co-occurrence statistics

Figure 4.4 presents the macro and micro F_1 scores achieved using *RF* as the collection frequency factor for terms and $\widehat{RF}$ for 2-termsets on Reuters-21578, 20 Newsgroups and OHSUMED where SVM is employed as the classification scheme. The terms selected using χ^2 are utilized as the BOW-based features and $\hat{\chi}^2$ defined in Eq. 3.1 is considered for 2-termset selection. The reference scores obtained using the baseline BOW-based representation are shown by the dashed lines. It can be seen in the figure that the 2-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 the BOW for large number of 2-termsets such as twice the number of terms used 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 a few hundred 2-termsets are employed on Reuters-21578 and 20 Newsgroups datasets whereas further improvements are achieved as the number of 2-termsets increases further on OHSUMED. This clearly shows that the number of discriminative 2-termsets is dataset dependent.

Using kNN, the macro and micro F_1 scores achieved on Reuters-21578, 20 Newsgroups and OHSUMED are presented in Figure 4.5. As in the case of SVM, the 2-termsets contribute to the scores on all three datasets. However, the highest scores are achieved using smaller numbers of features when compared to SVM. The performance drops that occur as the number of 2-termsets increases is mainly due to the inability of kNN to handle large feature spaces [18]. Comparing the performances of SVM and kNN, it can be seen that SVM provides superior scores than kNN in

both BOW-based and the proposed representations. Because of this, the experiments presented in the following context are conducted using only SVM.

Figure 4.6 presents the macro and micro F_1 scores achieved using $\widehat{RF}_{ind}$ for computing the termsubset weights. The F_1 scores obtained using $\widehat{RF}$ are also presented for comparison. The figures clearly demonstrate that the use of individual occurrences is fruitful on all three datasets. However, considering co-occurrences as well provides further improvements on Reuters-21578 and OHSUMED. Because of this, in the following context, $\widehat{RF}$ will be considered for termset weighting.

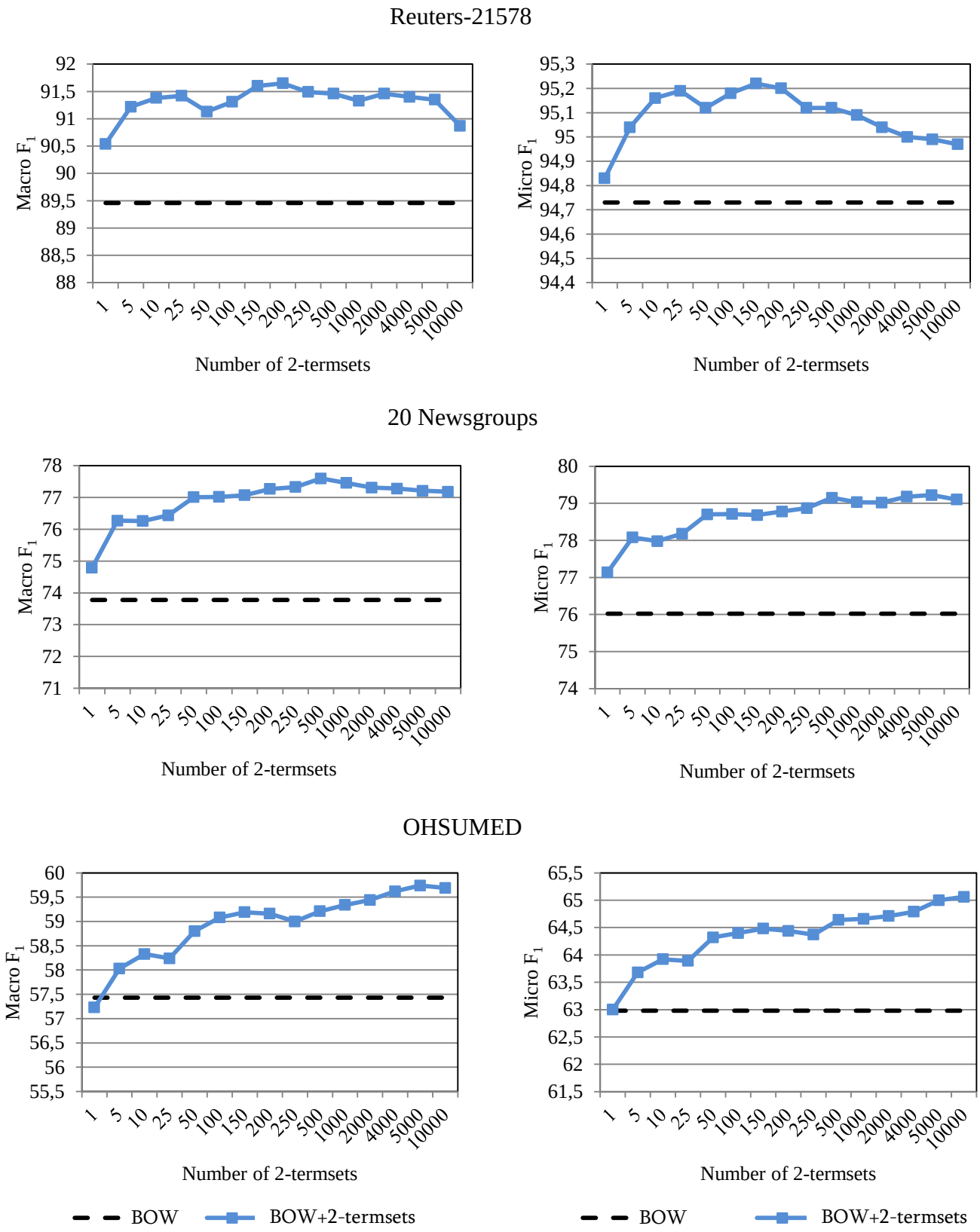


Figure 4.4: The macro and micro F_1 scores achieved by the proposed framework using *RF* and $\widehat{RF}$ as the collection frequency factors for the BOW-based features and 2-termsets respectively and SVM as the classification scheme.

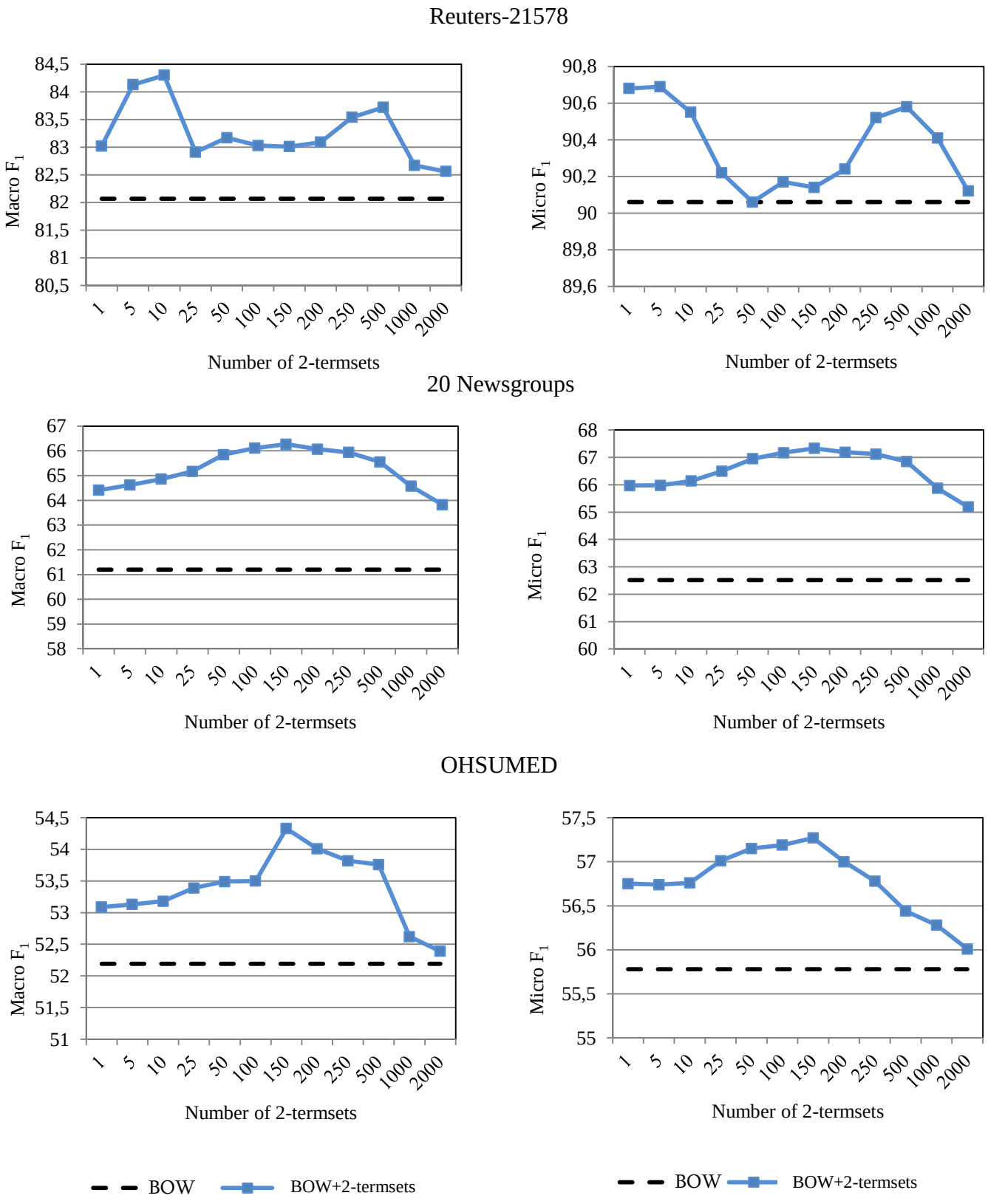


Figure 4.5: The macro and micro F_1 scores achieved by the proposed framework using RF and $\widehat{RF}$ as the collection frequency factors for the BOW-based features and 2-termsets respectively and kNN as the classification scheme.

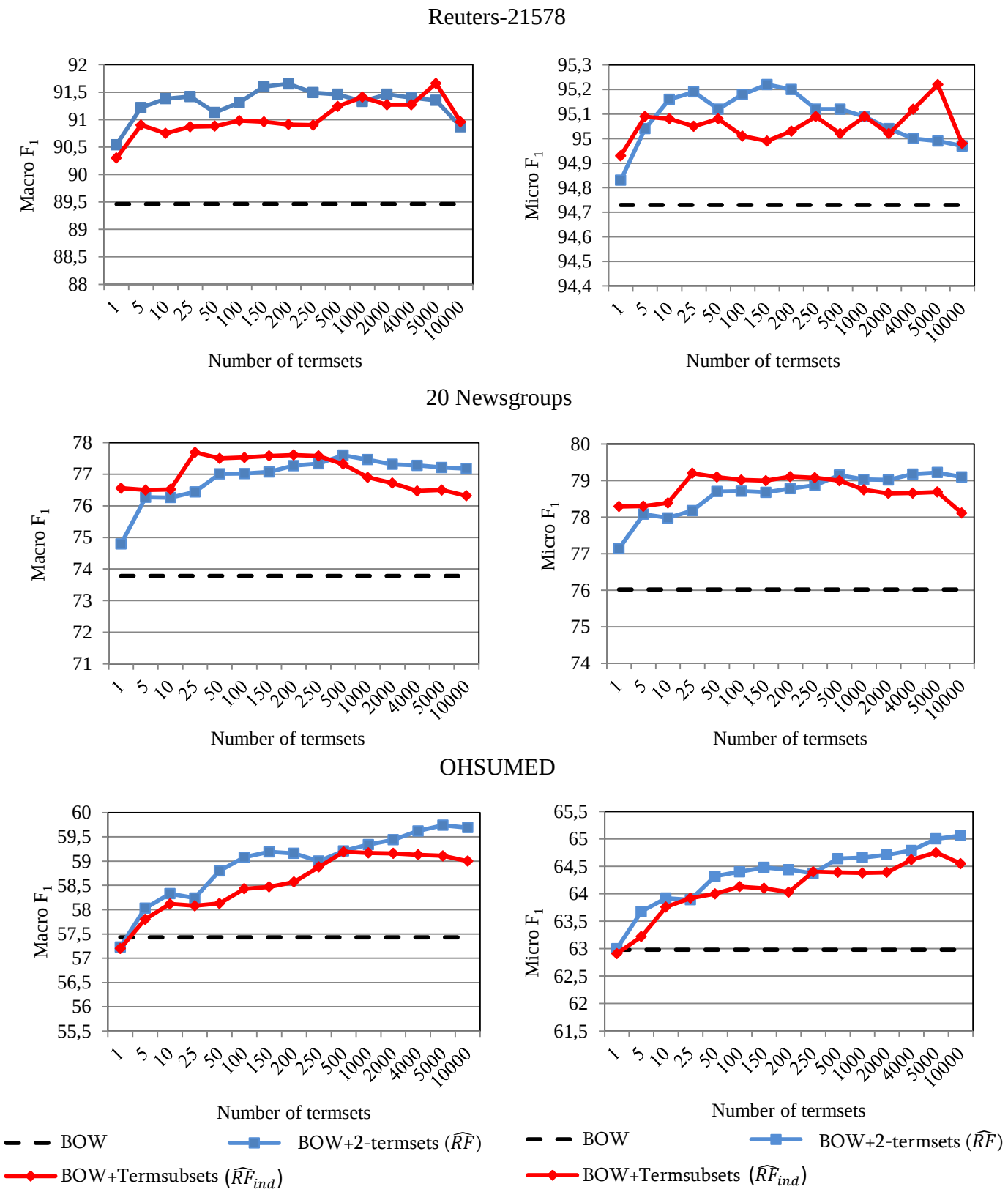


Figure 4.6: The macro and micro F_1 scores achieved by considering individual occurrences of terms but not their co-occurrence using $\widehat{RF}_{ind}$ as the collection frequency factor.

The experiments are repeated by using *MOR* and $\widehat{MOR}$ as the collection frequency factors. Figure 4.7 presents the macro and micro F_1 scores achieved, where the

BOW-based representation employing *MOR* as the collection frequency factor is also presented as a reference. It can be seen in the figure that improved 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 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 ($\frac{\hat{A}}{\hat{C}}$) are computed for each dataset over all categories. It should be noted that the value of this expression decreases as more termsets are selected from the negative class. Table 4.2 presents the values obtained using the top ranked 1000 2-termsets and the 2-termsets ranked between 9001 and 10000. It can be seen that the lower ranked 2-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, ($\frac{\hat{A}}{\hat{C}}$) < 1 means that the 2-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 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 2-termsets together with BOW are 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 25000 termsets for *MOR*. Both macro and micro F_1 scores slightly decrease for all three datasets when compared to 10000 termsets. In particular, the macro and micro F_1 scores are obtained as 90.26 and 94.89 for Reuters-21578, 72.69 and 74.88 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.

Table 4.2: The average ($\frac{\hat{A}}{\hat{C}}$) values obtained using the top ranked 1000 2-termsets and 2-termsets ranked between 9001 and 10000.

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

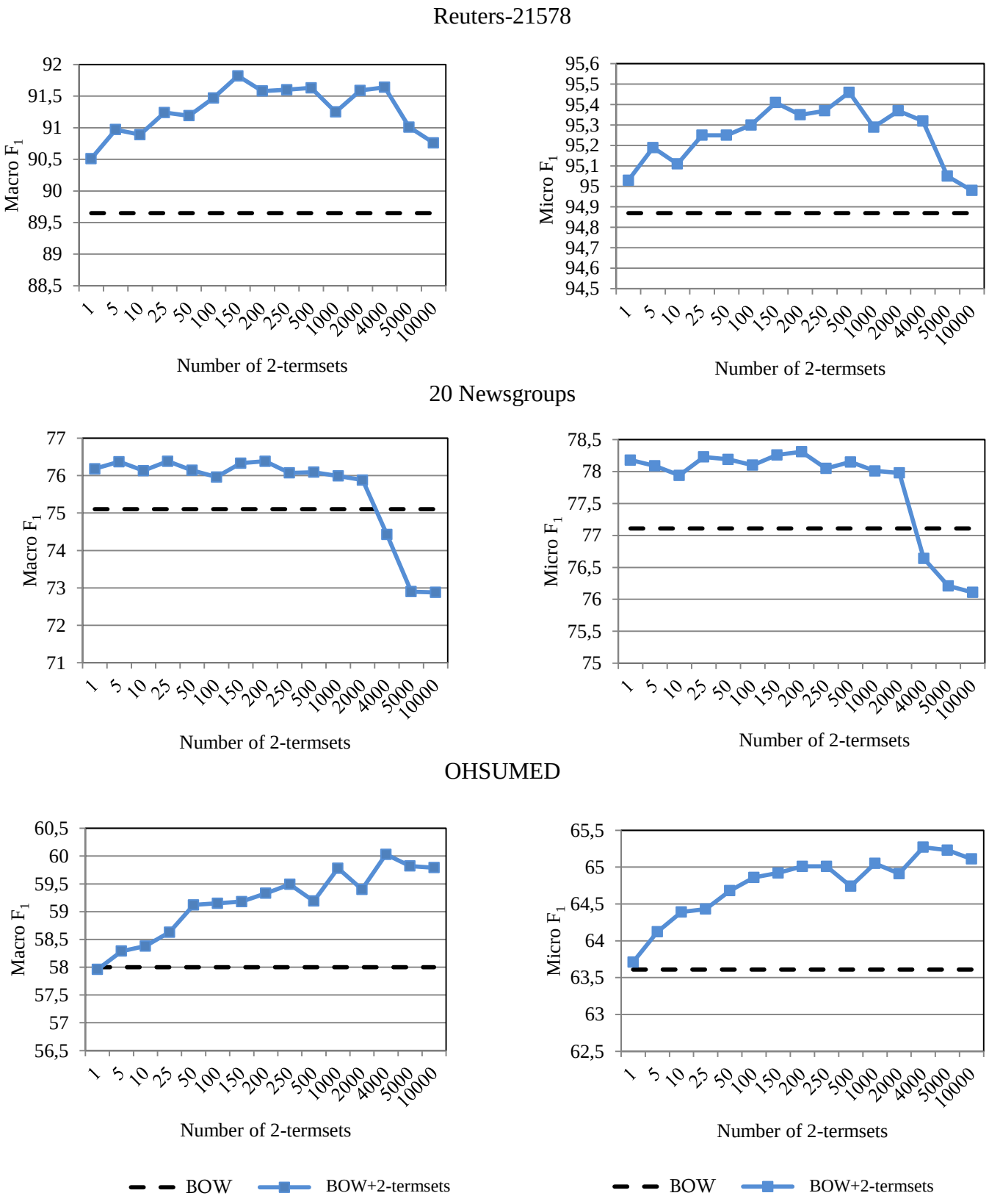


Figure 4.7: The macro and micro F1 scores achieved by the proposed framework using *MOR* and $\widehat{MOR}$ as the collection frequency factors for BOW and 2-termset based representations, respectively.

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 $\hat{\chi}^2$. Figure 4.8 presents the macro F_1 scores achieved by utilizing these schemes for 2-termset selection. *RF* and $\widehat{RF}$ are selected as the collection frequency factors for terms in BOW and 2-termsets, respectively. As it can be seen in 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 2-termsets selected by χ^2 and $\hat{\chi}^2$ are computed and presented in Figure 4.9. It can be seen in the figure that, on Reuters-21578, the termsets selected by χ^2 achieve higher $\hat{\chi}^2$ scores (around 1,000) 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 2-termsets in a better way than χ^2.

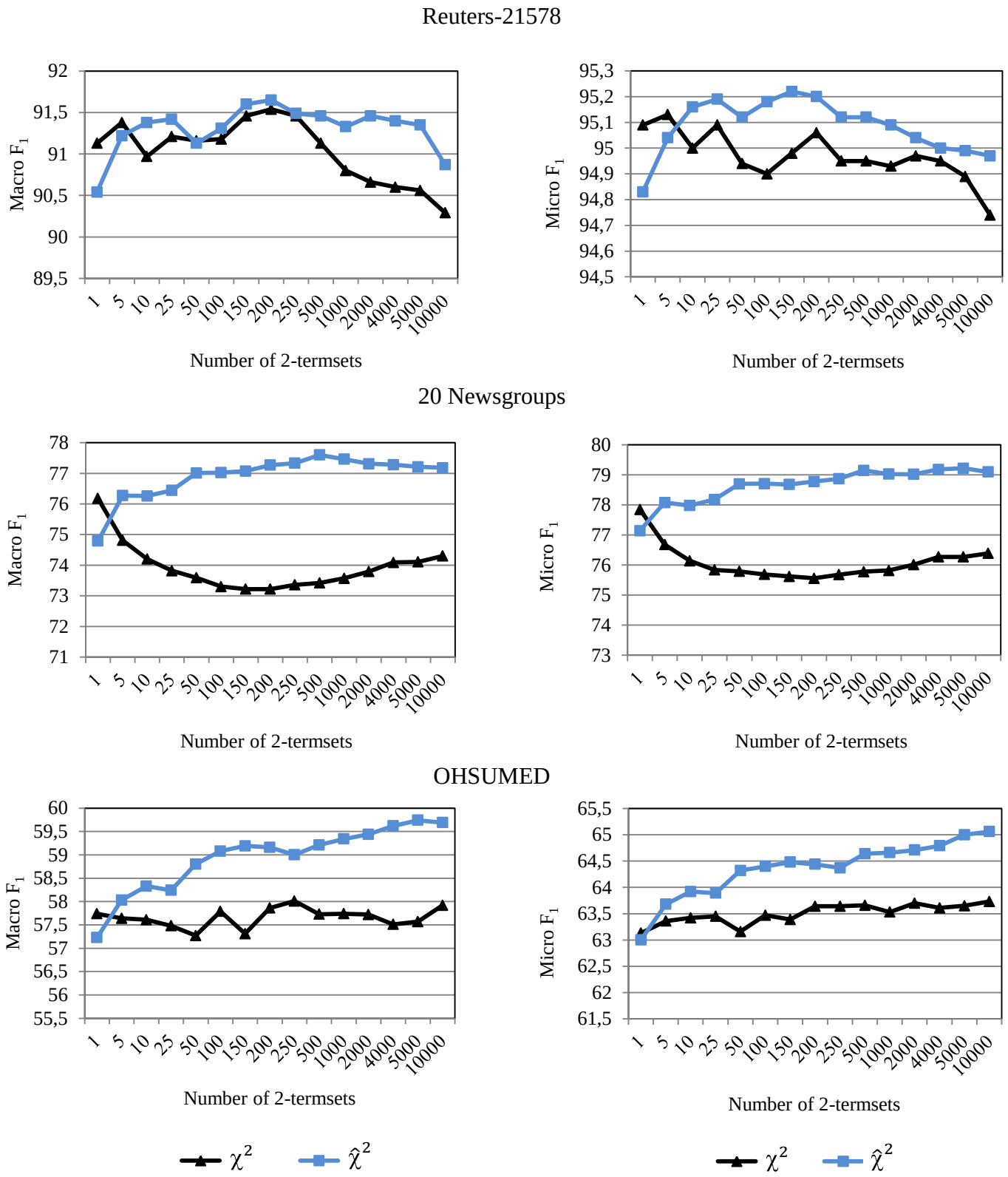


Figure 4.8: The macro and micro F1 scores achieved using χ^2 and $\hat{\chi}^2$ when *RF* and $\widehat{RF}$ are employed as the collection frequency factors for terms and 2-termsets, respectively.

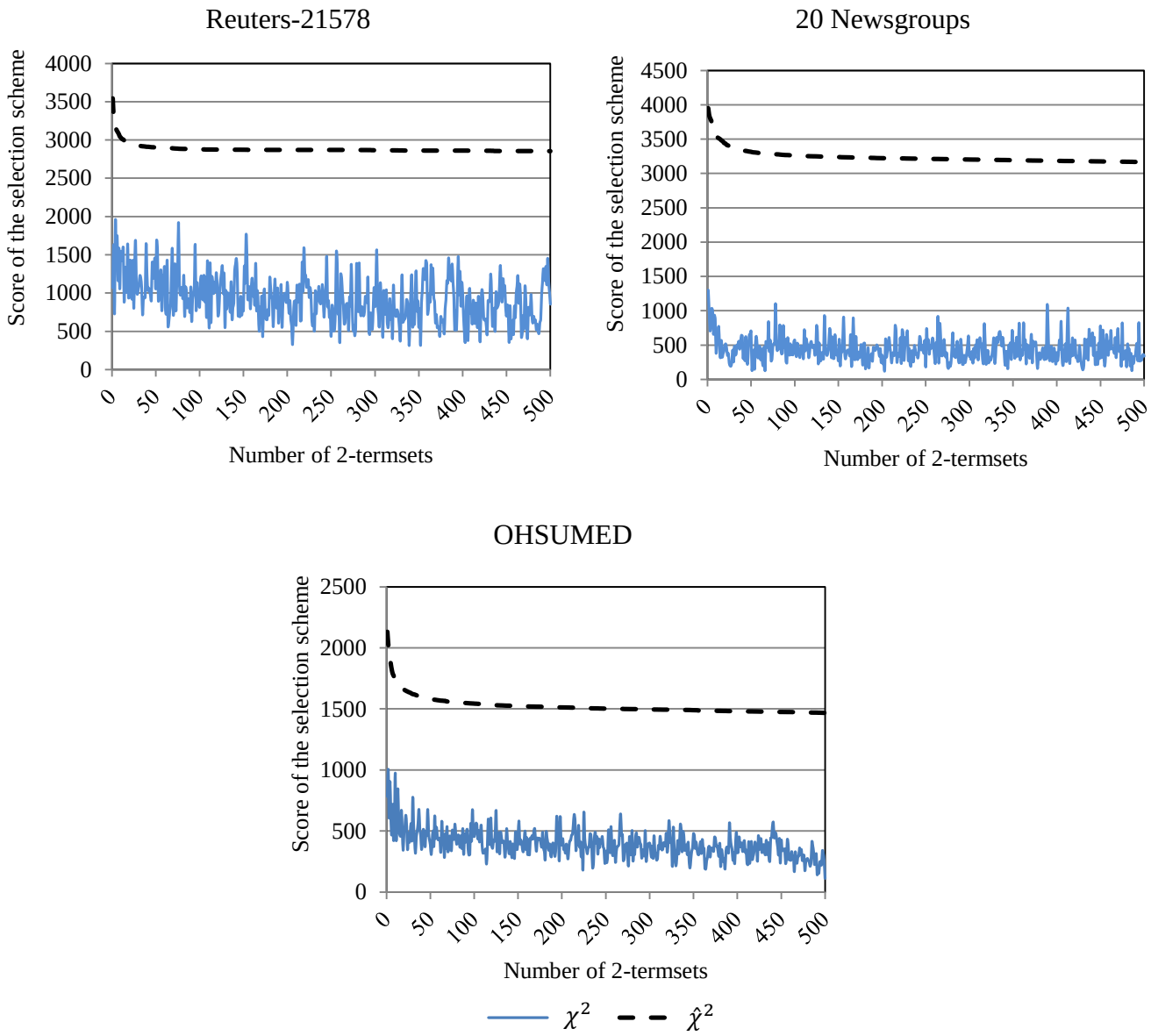
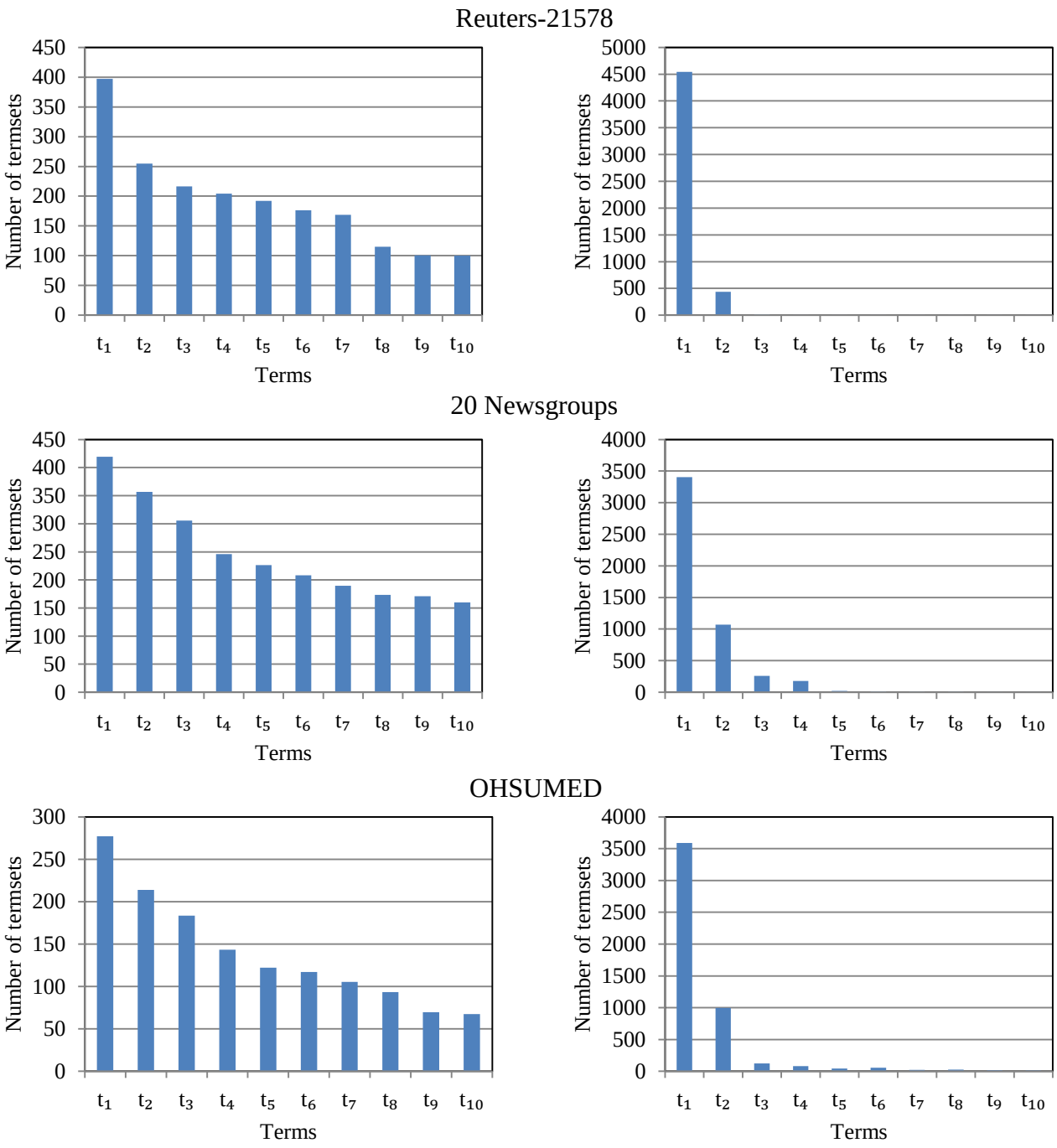


Figure 4.9: The $\hat{\chi}^2$ values of top 500 2-termsets selected by χ^2 and $\hat{\chi}^2$.

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. Figure 4.10 presents the average number of times that the most frequently used ten terms appear as members when 5000 2-termsets are employed. It can be seen in the figure that a small set of terms are members in a large number of 2-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 also be seen in the figure that the terms ranked fifth or above are used much

fewer times, and hence a corresponding bar does not even appear. On the other hand, in the case of χ^2, the most frequently used set of terms is larger. This means that χ^2employs a wider set of different terms as members in the 2-termsets.

The 2-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 2-termsets. Figure 4.11 presents the average number of different terms used in the 2-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 2-termsets. The rate decreases as the number of 2-termsets increases. For instance, on all datasets, approximately 500 different terms are employed in top ranked 500 2-termsets whereas, in the case of 5000 2-termsets, the number of different terms employed is approximately 3500 in 20 Newsgroups and OHSUMED.



(a) Using χ^2 for ranking

(b) Using $\hat{\chi}^2$ for ranking

Figure 4.10: The average number of times that the most frequently used ten terms appear as members when 5000 2-termsets are employed.

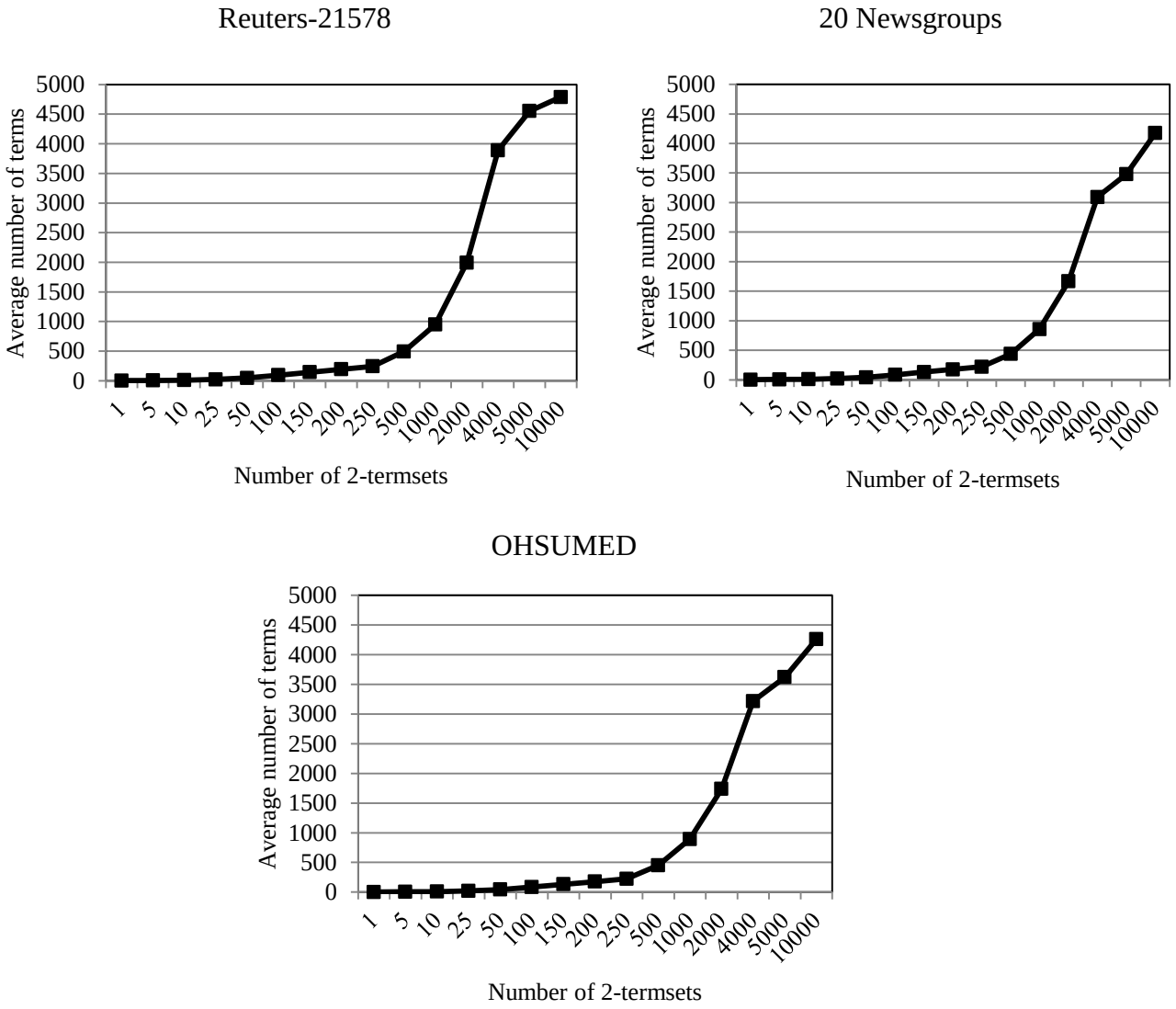


Figure 4.11: The average number of different terms employed in the 2-termsets selected using $\hat{\chi}^2$ as the termset selection scheme.

In our simulations, all 5000 terms utilized for BOW-based representation are considered for termset generation. This leads to (5000 × 4999)/2 2-termsets which is more than twelve million. 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 are also studied for termset generation. It should be noted that, for 500 terms, the total number of different termsets are reduced to be (500 × 498)/2 = 124750 which is a much smaller number. Figure 4.12 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 2-termset selection when 5000 terms are utilized. On a 3.1GHz i5 processor, the total number of minutes needed for computing and ranking the 2-termsets are computed as 38, 44 and 50 for the largest categories in Reuters-21578, 20 Newsgroups and OHSUMED, respectively.

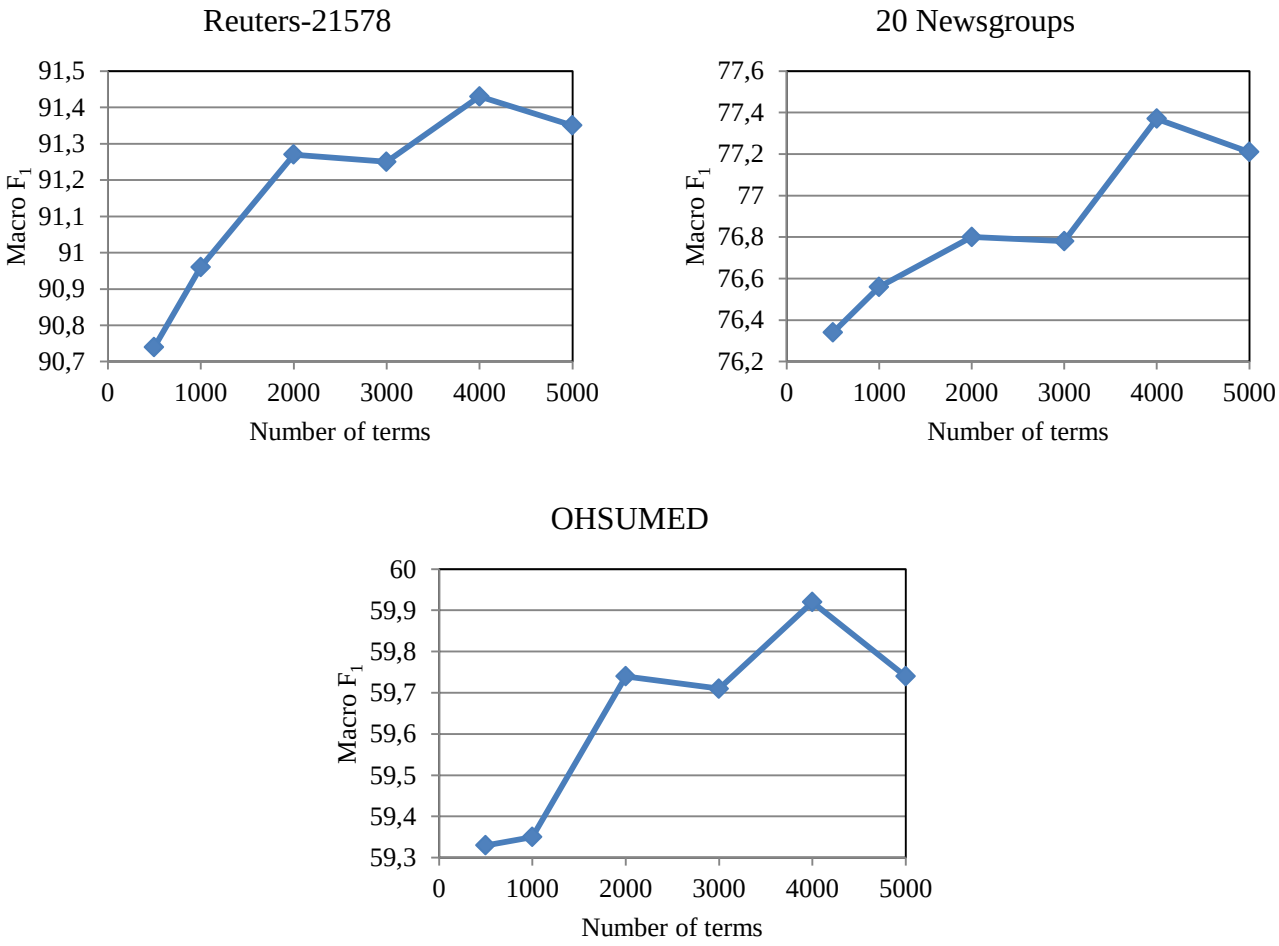


Figure 4.12: The macro F_1 scores achieved on three datasets using different number of terms for the 2-termset generation using RF and $\widehat{RF}$ as collection frequency factors.

As stated in Section 2.6, 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 2-termset selection. The results are presented in Figure 4.13.

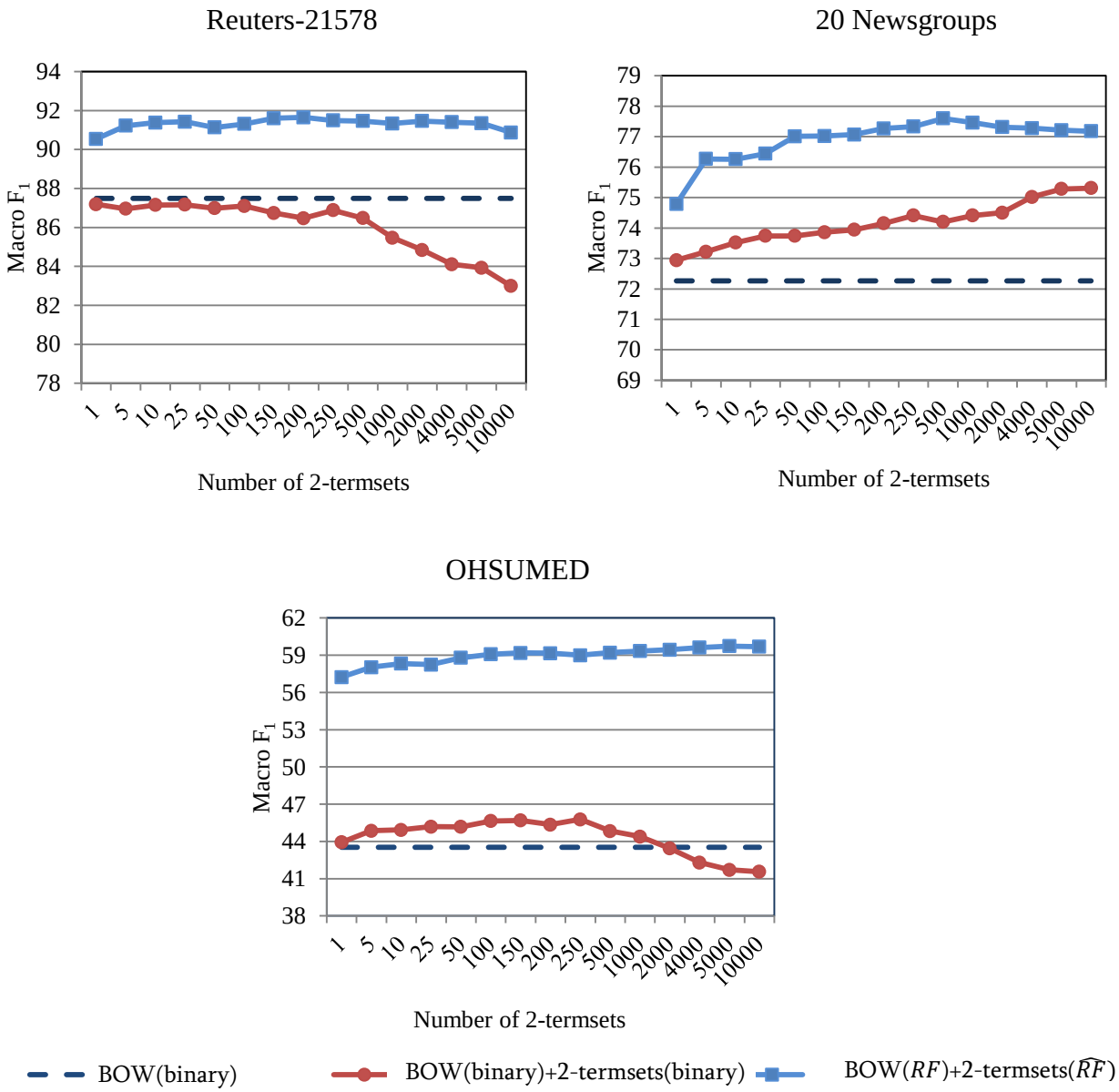


Figure 4.13: The macro F_1 scores achieved using χ^2 for both term and 2-termset selection. Binary term weighting is compared with $\widehat{RF}$.

The scores provided by the proposed approach using RF and $\widehat{RF}$ as the collection frequency factors and $\hat{\chi}^2$ for termset selection (denoted by BOW (RF)+2-termsets ($\widehat{RF}$)) are also presented for comparison. It can be seen in the figure that, when binary representation is employed for term weighting, the use of 2-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 three datasets.

In order to assess 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. Figure 4.14 presents the macro and micro F_1 scores achieved using RF as the collection frequency factor for the term weights and $\widehat{RF}$ for the 2-termset weights.

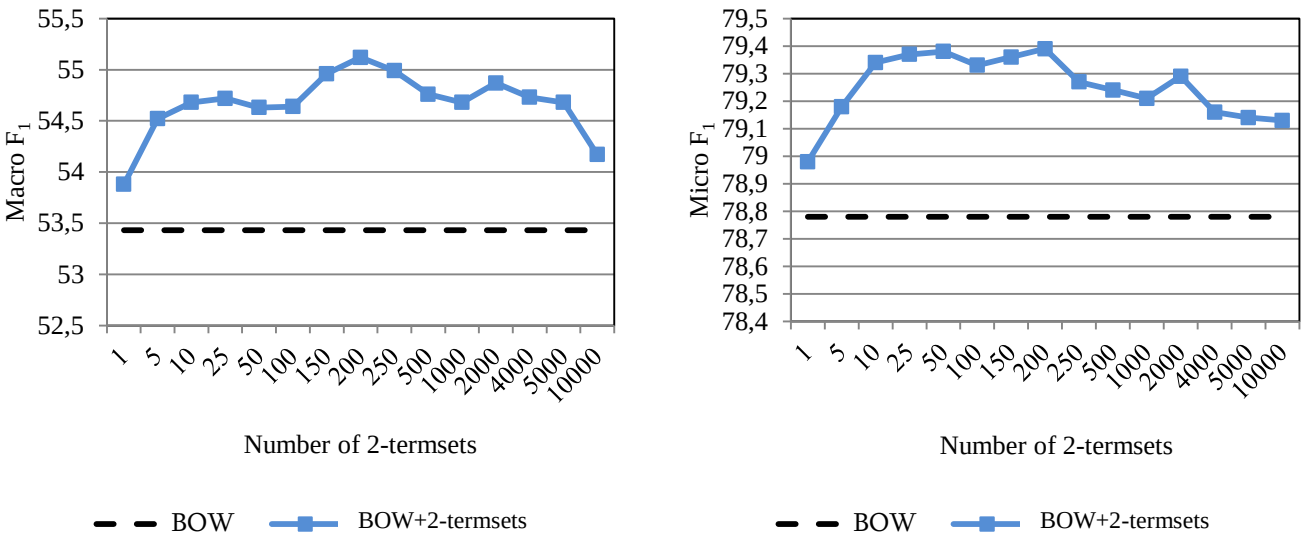


Figure 4.14: The macro and micro F_1 scores achieved on the entire Reuters collection by the proposed framework using *RF* and $\widehat{RF}$ as the collection frequency factors and SVM as the classification scheme.

Comparing Figures 4.4 and 4.14, it can be seen that consistent improvements are achieved also when less frequent categories are considered. Figure 4.15 presents the macro 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.

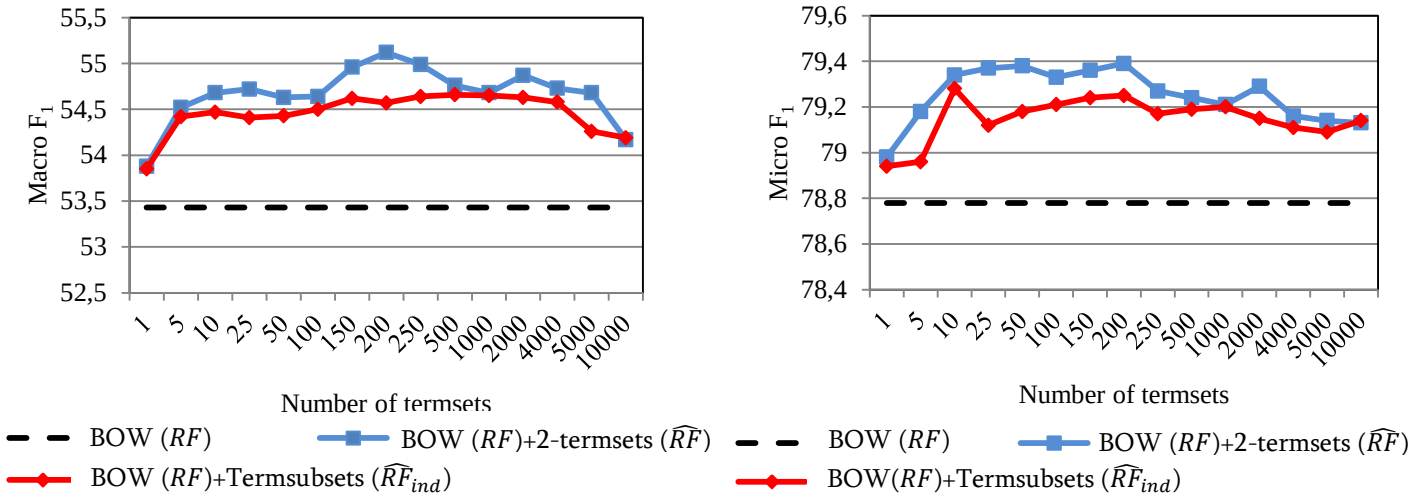


Figure 4.15: The macro and micro F_1 scores achieved on the entire Reuters collection by considering individual occurrences of terms without their co-occurrences using RF, $\widehat{RF}$ and $\widehat{RF}_{ind}$ as the collection frequency factors.

The relative performances of the selection schemes χ^2 and $\hat{\chi}^2$ are also in investigated on the entire Reuters collection. Figure 4.16 presents the macro F_1 scores achieved by utilizing these schemes for 2-termset selection. RF and $\widehat{RF}$ are selected as the collection frequency factors for terms and 2-termsets respectively. As it can be seen from the figures, better scores are achieved 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 Figure 4.7. However, larger differences are observed when less frequent categories are also considered.

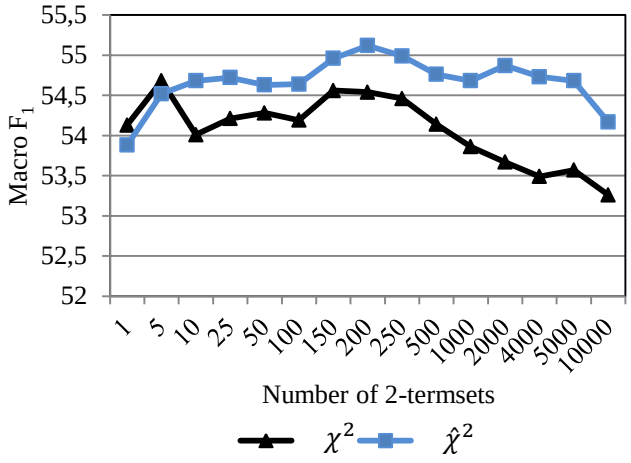


Figure 4.16: The relative performances of the selection schemes χ^2 and $\hat{\chi}^2$ on the entire Reuters collection when *RF* and $\widehat{RF}$ are employed as the collection frequency factors for terms and termsets respectively.

We compared the performance of the proposed framework with the binary representation for the entire Reuters corpus. The results are presented in Figure 4.17. 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 (*RF*)+2-termsets ($\widehat{RF}$)) are also presented for comparison. It can be seen in the figure that, when binary representation is employed for term weighting, the use of 2-termsets have only slight contribution 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.

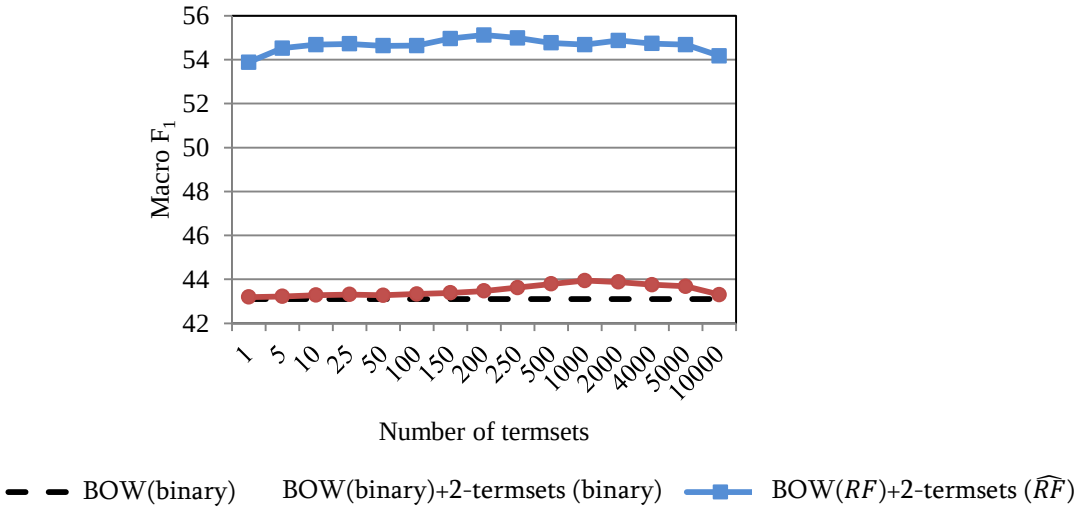


Figure 4.17: The macro F_1 scores achieved on the entire Reuters collection using χ^2 for both term and 2-termset selection and binary term weighting. The performance of the proposed scheme is also presented for reference where $\widehat{RF}$ is considered as the collection frequency factor.

4.4 Using co-occurrence statistics of bigrams

Weighting bigrams using co-occurrence statistics is also addressed. The Porter algorithm is firstly applied for stemming [21]. After computing all unigrams (individual terms) and bigrams, stop-words are eliminated from the lists of both unigrams and bigrams. Consequently, a bigram is not allowed to be made up of non-consecutive words that originally have a stop-word in between. All bigrams that include a stop-word are eliminated from the list.

After generating the lists of all unigrams and bigrams, we eliminate the bigrams that appear in less than three documents. Then, all terms are sorted using χ^2. The bigrams that include terms which are not in the top 5000 list are discarded. The remaining bigrams are then sorted using χ^2 being defined for bigrams.

The relative performances of $RF(\boldsymbol{b^2})$ and $RF'(\boldsymbol{b^2})$ are presented in Figure 4.18. It should be noted that, in computing $RF(\boldsymbol{b^2})$ of a bigram, *A* and *C* denote the positive and negative documents that include $\boldsymbol{b^2}$. The figure presents the macro and micro F_1 scores obtained by using RF for terms, $RF(\boldsymbol{b^2})$ and $RF'(\boldsymbol{b^2})$ for bigram weighting. The horizontal axis corresponds to the number of bigrams that are concatenated with 5000 terms. It can be seen in the figure that the performance increases as the number of bigrams is increased up to 4000 for 20 Newsgroups and OHSUMED datasets. It can also be seen that the proposed modification improves the performances on both 20 Newsgroups and OHSUMED. However the performance deteriorates for Reuters-21578. This means that, instead of considering and weighting only the co-occurrence of terms, the idea of considering the individual occurrences of the terms within the bigrams may be fruitful. However, the scores achieved are inferior to those obtained using 2-termsets as presented in Figure 4.4.

Binary weighting is generally considered as a reference when bigrams are employed. We compared the performance of the proposed scheme also with the binary representation. In particular, binary representation is used for both terms and bigrams. The results are presented in Figure 4.19. The results show that both macro and micro F_1 scores are improved on both 20 Newsgroups and OHSUMED when the number of bigrams employed is less than 500 and the scores are degrade for Reuters-21578. On OHSUMED dataset, the scores drop below the baseline system when the number of bigrams is increased above 500. However, the scores achieved are far below those that are achieved by the proposed framework as presented in Figure 4.4.

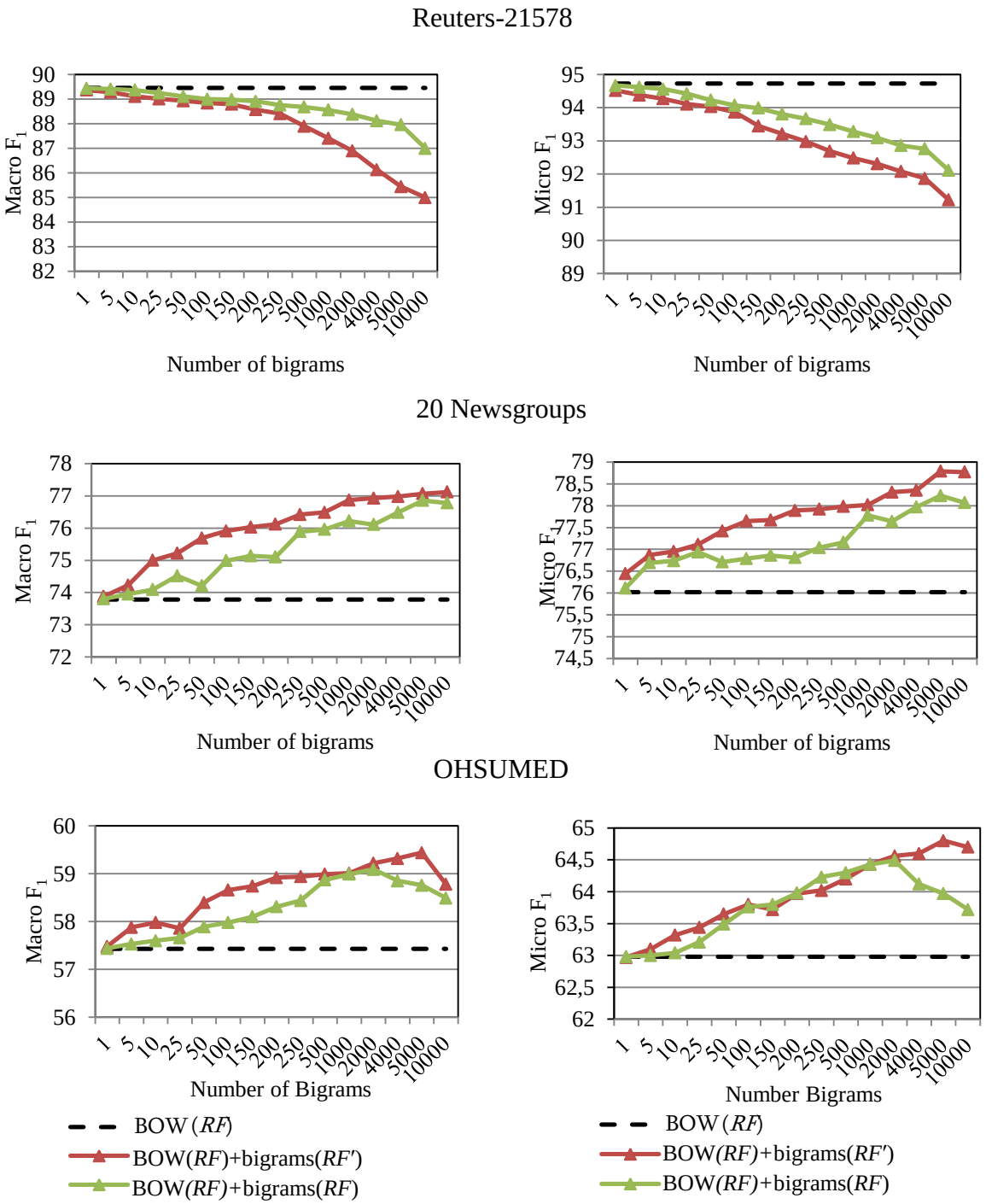


Figure 4.18: The macro and micro F_1 scores achieved using $RF(\boldsymbol{b^2})$ and $RF'(\boldsymbol{b^2})$ as the collection frequency factors.

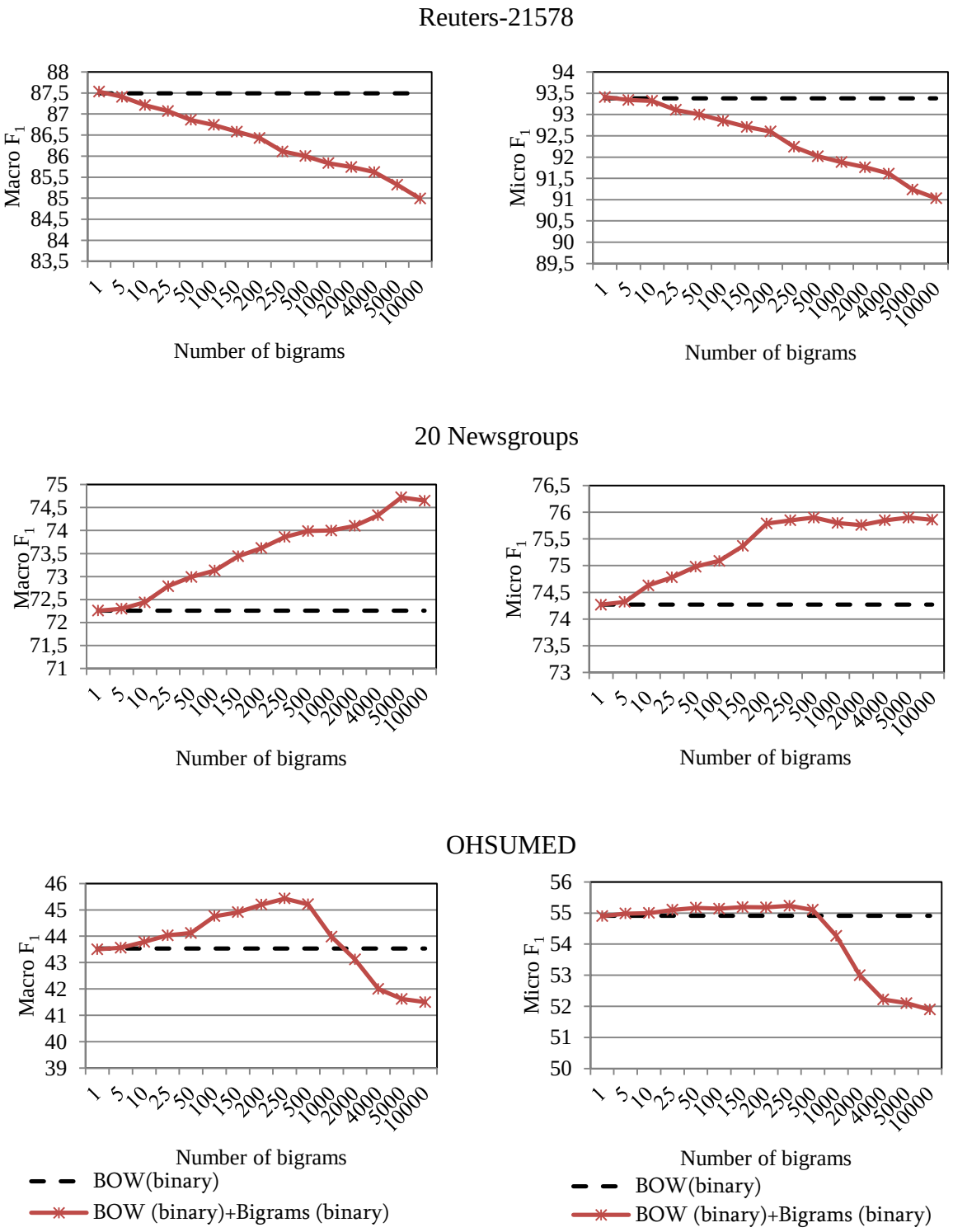


Figure 4.19: The macro and micro F_1 scores achieved using the binary representation for both terms and bigrams.

4.5 Using cardinality statistics for 2-termsets

The comparison between the use of co-occurrence statistics and the cardinality statistics based weighting for 2-termsets is shown in Figure 4.20. The figure presents

the macro and micro F_1 scores achieved using RF, $\widehat{RF}$ and $\widetilde{RF}$ as the collection frequency factors. The performance of BOW-based is also presented for comparison. It can be seen in the figure that $\widehat{RF}$ provides better scores compared to $\widetilde{RF}$ on all three datasets. Because of this, when studying the effectiveness of 3-termsets, $\widehat{RF}$ will be employed for the 2-termsets.

4.6 Using co-occurrence statistics of 3-termsets

The use of 3-termsets together with terms and 2-termsets is also studied in this thesis. It can be seen in Figure 4.4 that when all three datasets are considered, the best scores are obtained when 500 2-termsets are used. Because of this, the number of 2-termsets is set to be 500 for all three datasets.

Figure 4.21 shows the macro and micro F_1 scores achieved using $\widehat{RF}$ as the collection frequency factor for both 2-termsets and 3-termsets. The results corresponding to BOW(RF)+2-termsets ($\widehat{RF}$) using 500 2-termsets are also presented for comparison. It can be seen in the figures that the use of the 3-termsets contributes to the performance when small number of 3-termsets is employed. It can be concluded that the statistical information about the co-occurrences may not be reliably estimated for a large number of termsets as the length of the termsets increase.

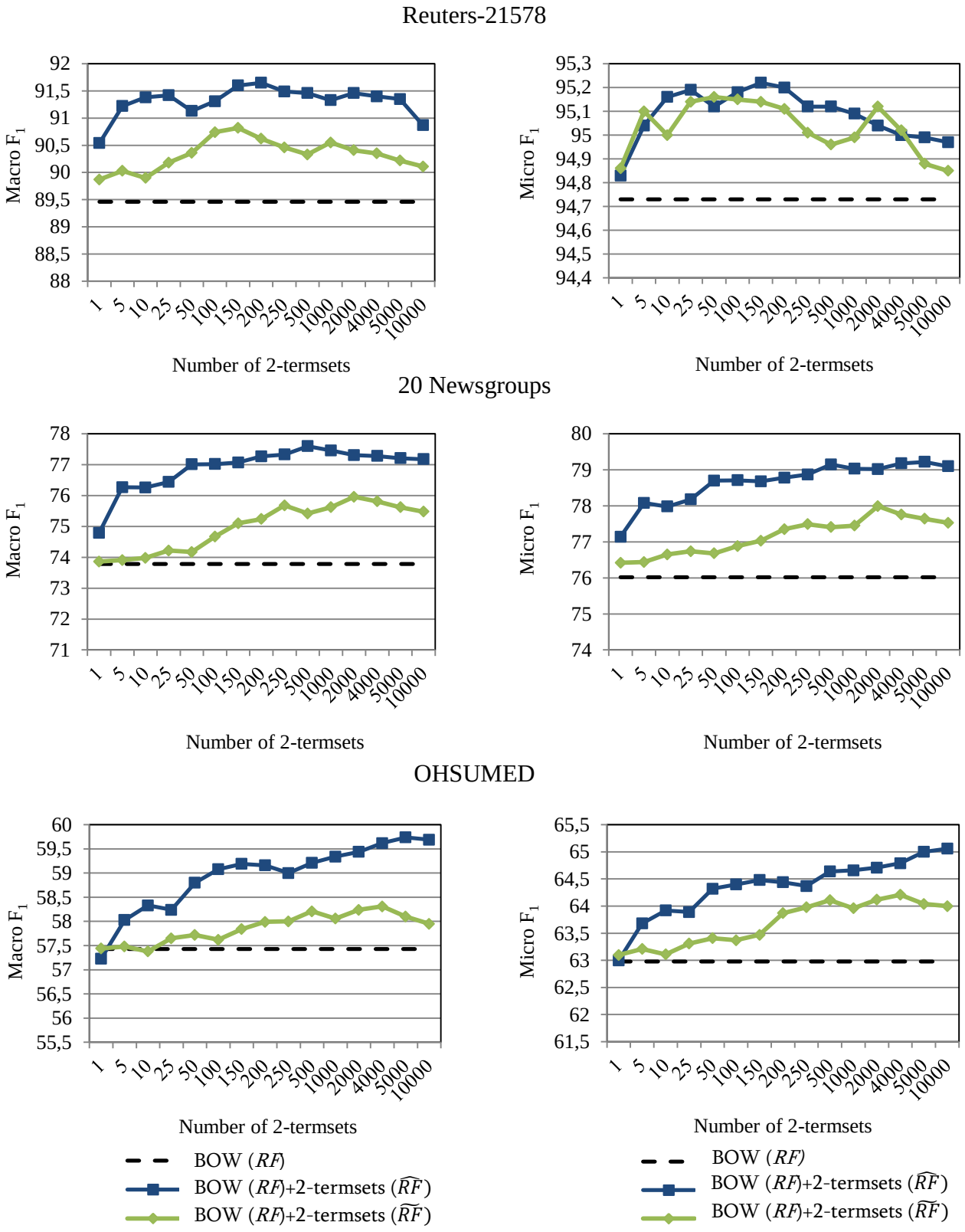
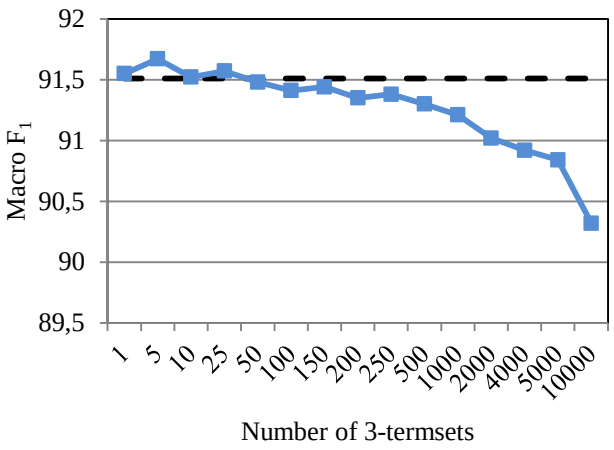
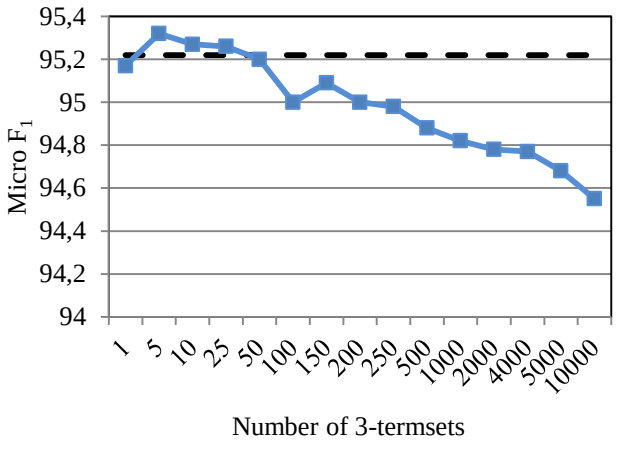


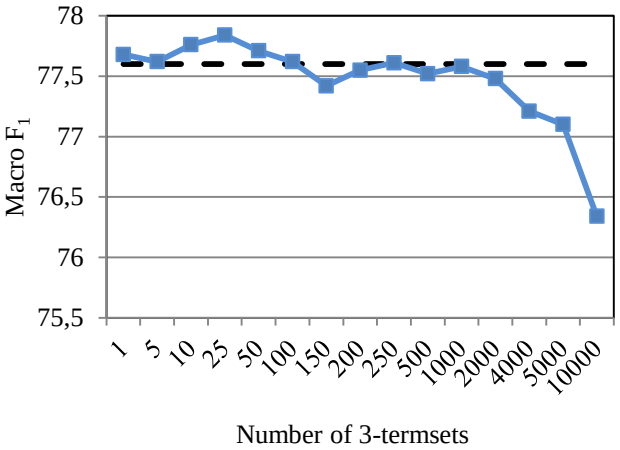
Figure 4.20: The macro and micro F_1 scores achieved using RF, $\widehat{RF}$ and $\widetilde{RF}$ as the collection frequency factors.

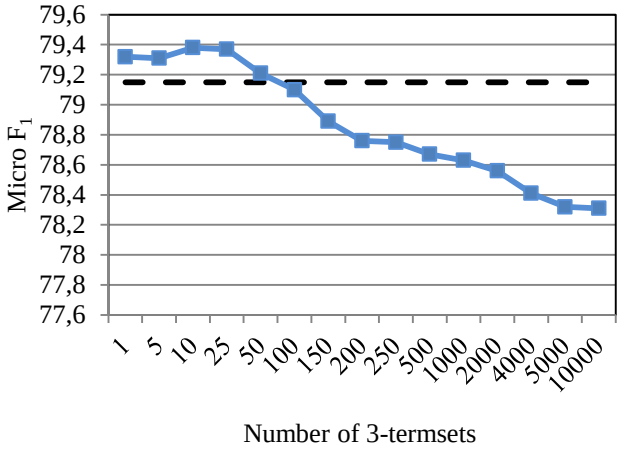
Reuters-21578



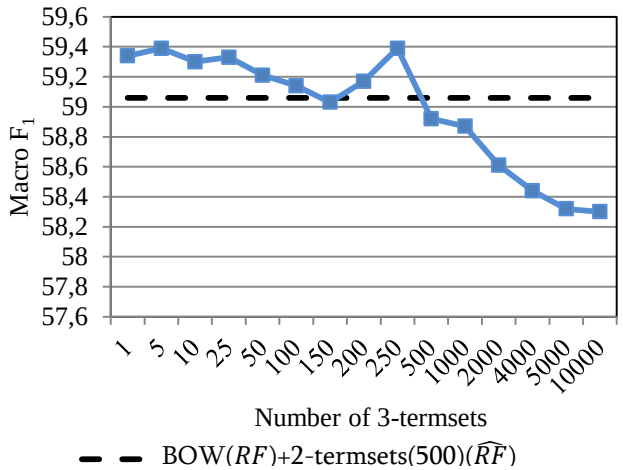


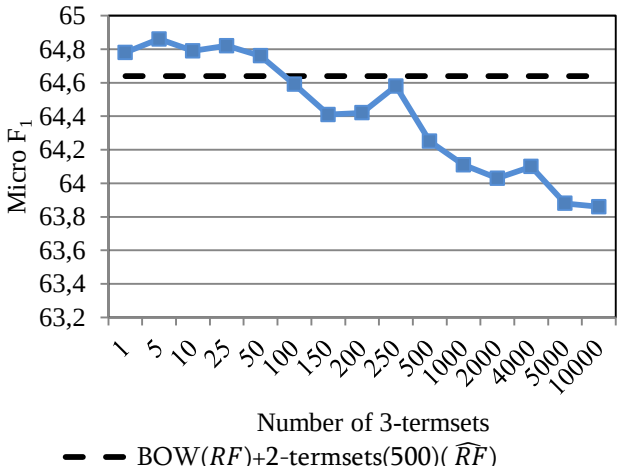
20 Newsgroups





OHSUMED





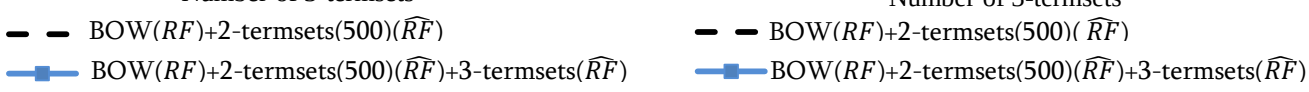


Figure 4.21: The macro and micro F_1 scores achieved by the proposed framework using 3-termsets.

4.7 Using cardinality statistics for 3-termsets

As mentioned in Chapter 2, reliable estimation of cardinality statistics may not be possible as the length of the termsets increase. However, this is less likely to occur when only the cardinalities are considered. In order to investigate this, the use of $\widetilde{RF}$ for weighting the 3-termsets is addressed. Figure 4.22 shows that the scores obtained by using $\widetilde{RF}$ for the 3-termsets are superior to the scores achieved by $\widehat{RF}$ when the number of 3-termsets is large . For instance, when 500 3-termsets are considered, superior macro and micro F_1 scores achieved by using $\widetilde{RF}$.

Top 50 3-termsets obtained for the largest and smallest categories are shown in Appendix A, Appendix B and Appendix C for Reuters-21578, 20 Newsgroups and OHSUMED, respectively. It can be seen in the tables that the top termsets include individually discriminative terms which are the top ranked terms such as "corn" in corn category and "ct" in earn category of Reuters-21578.

4.8 Using cardinality statistics for 4-termsets

The scores achieved using the cardinality statistics for 4-termsets are presented in Figure 4.23. It can be seen in the figure that the use of 4-termset provides a more robust representation, leading to improved scores that are more distinctive on OHSUMED.

In order to identify the 4-termsets that contribute to the performance, top 10 4-termsets computed for the largest and smallest categories are presented in Table 4.3, 4.4 and 4.5 for Reuters-21578, 20 Newsgroups and OHSUMED, respectively.

When all datasets are considered, it can be argued that the 4-termsets include individually relevant terms and domain-specific information is necessary to comment on the importance of co-occurrence of the terms in these termsets.

Table 4.3: Top 10 4-termsets obtained for the categories earn and corn of Reuters-21578.

4-termsets	earn	corn
1	ct, tax, respons, director	corn, soybean, agriculture, rebat
2	ct, announc, soviet, chanc	corn, maiz, wheat, depart
3	shr, agreement, mth, rule	corn, rebat, pik, Bueno
4	net, rev, jan, sell	maiz, belt, subsidi, program
5	profit, washington, today, stabilis	corn, kansa, agrianalysi, unpublish
6	div, ad, iran, origin	tonn, grower, depart, rebat
7	rev, bank, unit, rule	agriculture, unknown, deliveri, cordoba
8	purchas, good, link, compet	usda, feed, licenc, total
9	dividend, record, pai, econom	tonn, harvest, reduc, evnsvll
10	qtly, export, week, discuss	ec, dry, barg, delink

Table 4.4: Top 10 4-termsets obtained for the categories alt.atheism and talk.religion.misc of 20 Newsgroups.

4-termsets	alt.atheism	talk.religion.misc
1	relig, pathogen, galacticent, exchang	misc, scrub, spoil, weinss
2	overgraze, exchang, fondli, pragmat	misc, scrub, milton, psalm
3	meantime, pragmat, crumenam, exchang	spoil, psalm, welfare, taint
4	ozguven, repercuss, profit, minnestoa	weinss, prettier, bandwagon, anthro
5	minnestoa, pittsburgh, motorola, willingli	anthro, dsav, airwai, disnei
6	motorola , wilaya, utdalla, utopia	prettier, airwai, deceas, abolit
7	pathogen,pittsburgh, utdalla, vote	abolit, heat, dealt, illumin
8	tianiti, undoubt, uneasi, Abraham	dealt, flamabl, magu, bahama
9	undoubt, uneasi, campaign, campbel	bahama, abolit, irag, ceas
10	galacticent, pragmat, hume, aspirin	misc, psalm, disord, widengren

Table 4.5: Top 10 4-termsets obtained for the categories Bacterial Infections and Mycoses and Pathological Conditions, Signs and Symptoms of OHSUMED.

4-termsets	Bacterial Infections and Mycoses	Pathological Conditions, Signs and Symptoms
1	bacteri, abnorm, agent, alloxan	patholog, symptom, paracarcin, iranian
2	nodal, abdomin, biomechan, ofloxacin	massach, intercartilag, revert, overantic
3	abandon, postradiotherapi, actinomycin, contagiosum	intercartilagen, outward, patient, formul
4	posttransplant, abandon, abbott, accomplish	massach, inspir, flecainid, flunitrazepam
5	abbott, gradient, grandmoth, griseofulvin	patholog, ipth ipth, cordocentesi
6	disturb, diurnal, domest, ecmo	flecainid,constraint, conjunctiv gynaecolog
7	bacteri, lithotripsi, hyperammonem, achalasia	symptom, gynaecolog, microfollicl, exoplasm
8	lithotripsy, birmingham, juxtaren, karnofski	ethanol, doctor, cholesterol, fibrointim
9	paraganglioma, abandon, saccharomyc, tamoxifen	gynaecolog, microfollicl, extraembryon, rhombic
10	tamoxifen, extraembryon, diseas, disciplinary	symptom, uret, herniotomi, acadian

4.9 Summary of the experimental results

Table 4.5 presents the summary of the results achieved using the proposed scheme and the baseline. It can be seen in the figures that the best F_1 scores typed in boldface are achieved when ten 4-termsets are used in addition to 5,000 terms, 500 2-termsets and 500 3-termsets in majority of the cases.

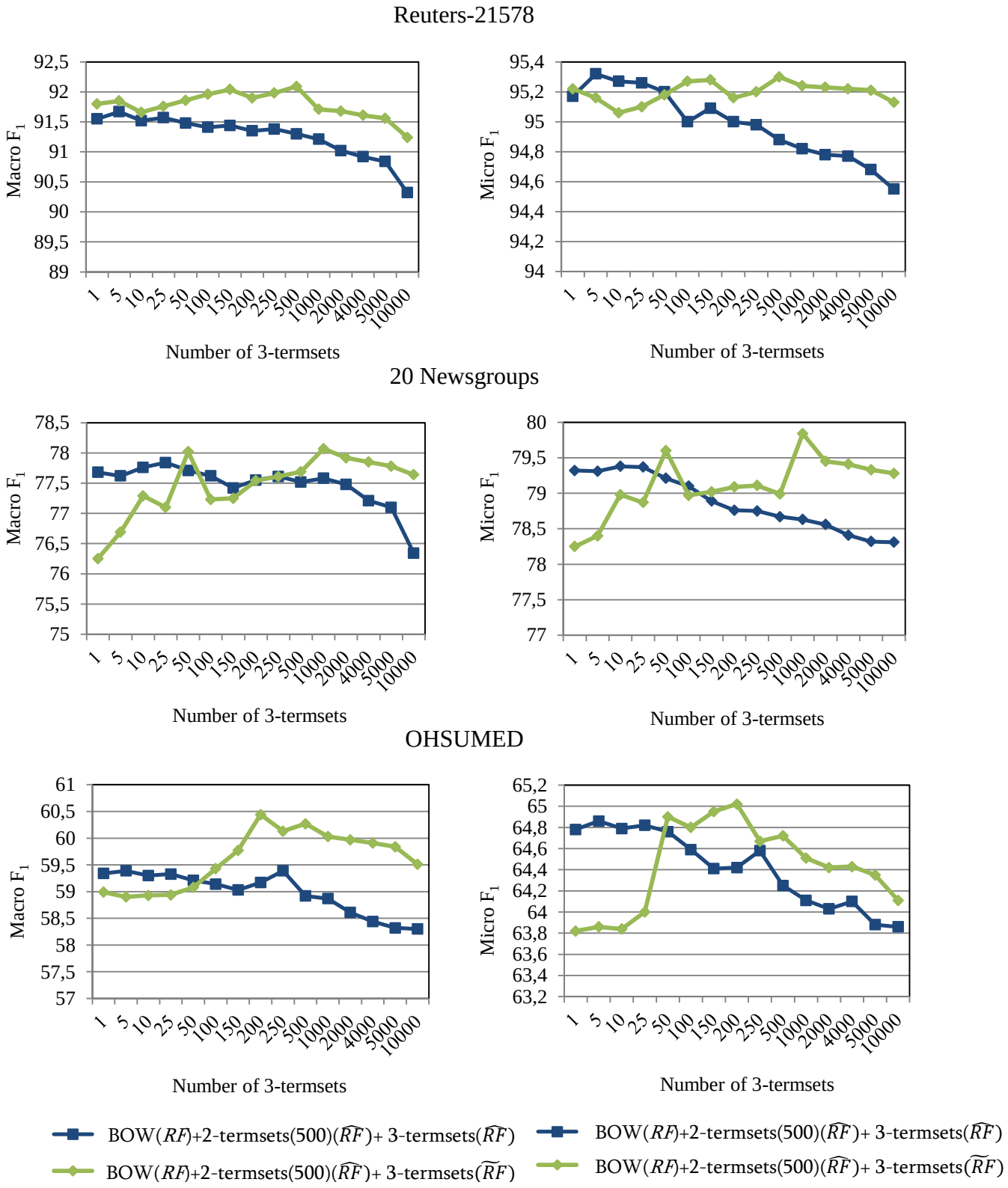
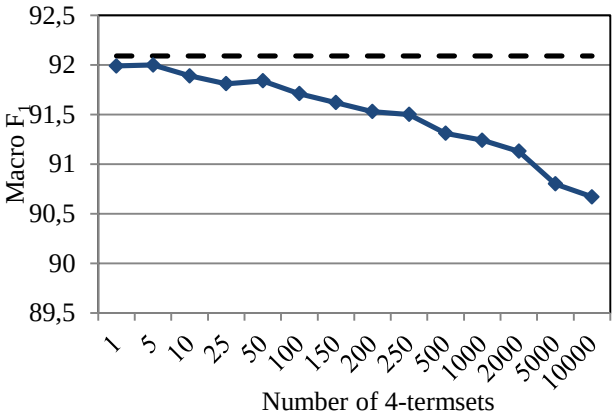
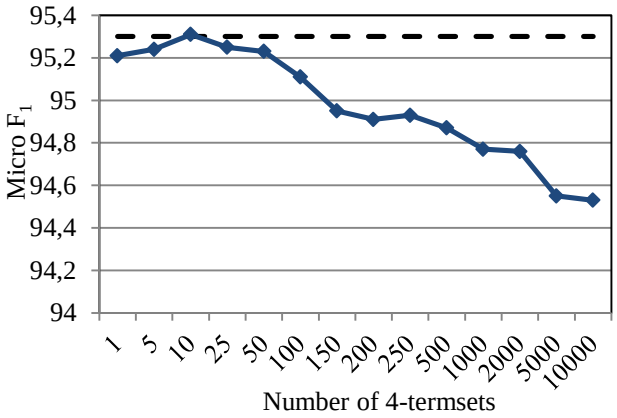


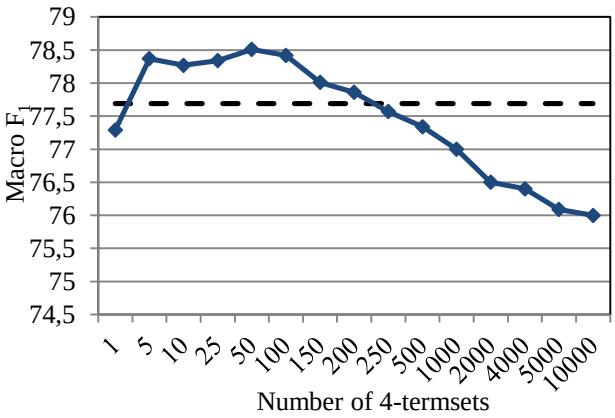
Figure 4.22: The macro and micro F_1 scores achieved using 3-termsets.

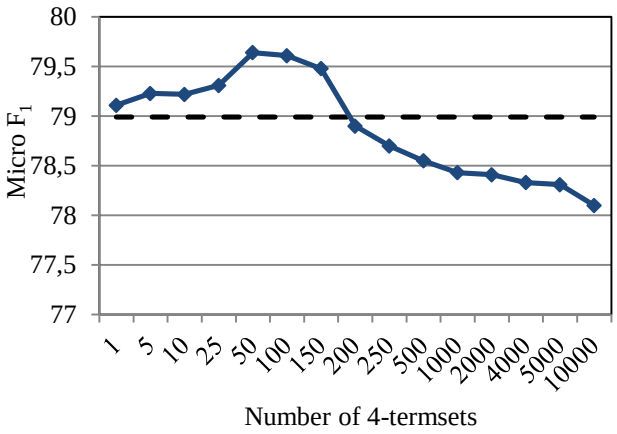
Reuters-21578



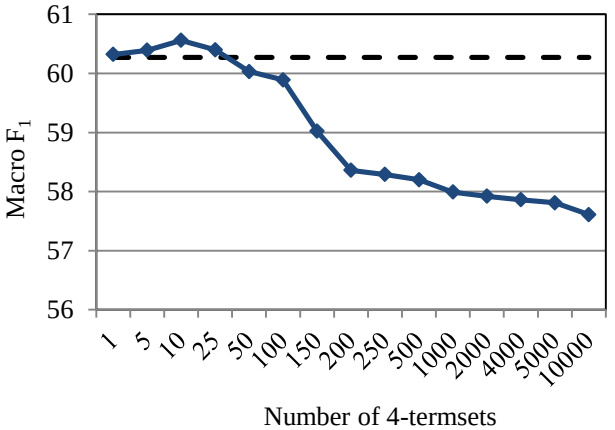


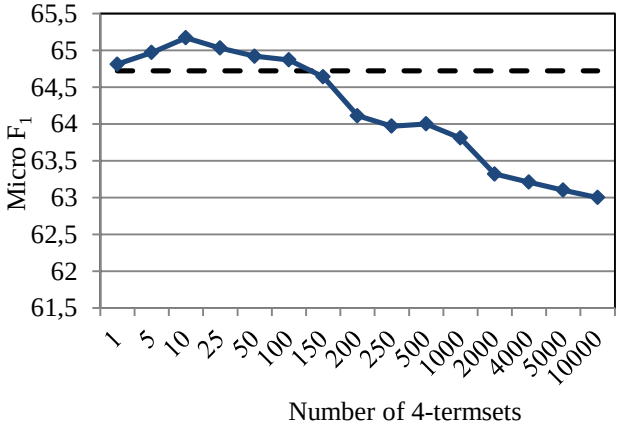
20 Newsgroups





OHSUMED





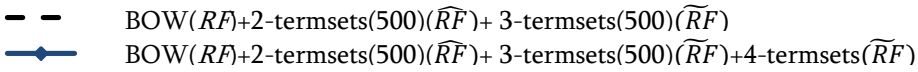


Figure 4.23: The macro and micro F_1 scores achieved using 4-termsets.

Table 4.6: The macro and micro F_1 scores obtained using the proposed scheme and the baseline.

Categorization System	Reuters-21578		20 Newsgroups		OHSUMED	
	macro F_1	micro F_1	macro F_1	micro F_1	macro F_1	micro F_1
Baseline ($tf \times RF$)	89.46	94.73	73.78	76.02	57.43	62.98
BOW(RF) + 2-termsets(500)($\widehat{RF}$)	91.46	95.12	77.6	79.15	59.21	64.64
BOW(RF) + 2-termsets(500)($\widetilde{RF}$)	90.33	94.96	75.42	77.41	58.21	64.11
BOW(RF) + 2-termsets(500)($\widehat{RF}$) + 3-termsets(500)($\widetilde{RF}$)	**92.09**	95.30	77.69	78.99	60.27	64.72
BOW(RF) + 2-termsets(500)($\widehat{RF}$) + 3-termsets(500)($\widetilde{RF}$)+ 4-termsets(10)($\widetilde{RF}$)	91.89	**95.31**	**78.27**	**79.22**	**60.56**	**65.17**

Chapter 5

CONCLUSION AND FUTURE WORK

In this dissertation, a novel framework is proposed for selecting and weighting termsets. The definition of termset-based features is revised where the joint occurrence statistics of the terms are utilized for termset selection and weighting. This allowed a termset to be assigned a nonzero weight even if all member terms do not appear in the document under concern. The main motivation for this approach is explained by an example.

The joint occurrences of the individual terms within 2-termsets including two terms is firstly 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 provided remarkably superior macro and micro F_1 scores compared to the baseline that employs BOW representation. The proposed approach for termset selection scheme is also compared with the conventional selection schemes. More specifically, 2-termset selection using χ^2 and its adapted form are compared where consistent improvements are observed on three benchmark datasets.

The proposed framework is then extended to employ both 2-termsets and 3-termsets to enrich the BOW-based representation. The experiments have shown that, when 3-termsets are used together with 2-termsets, better scores are achieved when

compared to employing BOW and 2-termsets only. However, superior scores are achieved only when small number of 3-termsets (50 or less) is considered and the performance is observed to degrade when more 3-termsets are used. It is emphasized that the statistical information about the co-occurrences may not be reliably estimated as the length of the termsets increase.

As a solution to this problem, employing the cardinality statistics of termsets for their weighting is addressed. It is observed that, although the use of cardinality statistics provides inferior scores compared to co-occurrence statistics for 2-termsets, the use of cardinality statistics based weights lead to better scores when 3-termsets and 4-termsets are employed.

In order to evaluate the proposed weighting scheme, weighting bigrams is also addressed. In these experiments, non-zero weights are assigned to bigrams even if only one of the terms occurs. The co-occurrence statistics of the terms that constitute bigrams is studied to develop a better weighting scheme. Experiments conducted have shown that the proposed scheme contributes to the performance of BOW-based representation in two datasets and degrades for third. However, the scores are observed to be inferior on all three datasets when compared to the use of 2-termsets.

As a future research, selection of the best-fitting subset of terms and termsets of different lengths should be addressed. Moreover, in a recent study, it is shown that the use of term frequencies may boost the performance of term selection schemes. This observation should be investigated for selecting better termsets.

REFERENCES

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

[2] Chen, J., Huang, H., Tian, S., and Qu, Y. (2009). Feature selection for text classification with naïve Bayes. Expert Systems with Applications, 36: pp. 5432–5435.

[3] Liu, Y., Loh, H. T., and Sun, A. (2009). Imbalanced text classification: A term weighting approach. Expert Systems with Applications, 36: pp. 690–701.

[4] Yang, J., Liu, Y., Zhu, X., Liu, Z., and Zhang, X. (2012). A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization. Information Processing Management, 48(4): pp. 741–754.

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

[6] Mladenic, D. and Grobelnik, M. (1998). Word sequences as features in text-learning. In Proceedings of the 17th Electrotechnical and Computer Science Conference (ERK98), pp. 145–148.

[7] Lewis, D. D. (1992b). Representation and learning in information retrieval. PhD thesis, Amherst, MA, USA. UMI Order No. GAX92-19460.

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

[9] Tan, C. M., Wang, Y. F., and Lee, C. D. (2002). The use of bigrams to enhance text categorization. In Information Processing Management, 38: pp. 529–546.

[10] Bekkerman, R. and Allan, J. (2004). Using bigrams in text categorization. Technical Report IR-408, Center of Intelligent Information Retrieval, UMass Amherst.

[11] Figueiredo, F., Rocha, L., Couto, T., Salles, T., Gon¸calves, M. A., and Meira, W. (2011). Word co-occurrence features for text classification. Information Systems, 36(5): pp. 843–858.

[12] Tesar, R., Poesio, M., Strnad, V., and 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, pp. 138–146, New York, NY, USA. ACM.

[13] 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, pp. 37–50, New York, NY, USA. ACM.

[14] Boulis, C. and 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.

[15] Özgur, L. and Güngör, T. (2010). Text classification with the support of pruned dependency patterns. Pattern Recognition Letters, 31(12): pp. 1598–1607.

[16] 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.

[17] Baker, D. L. and 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, pp. 96–103, New York, NY, USA. ACM.

[18] Lan, M., Tan, C. L., Su, J., and Lu, Y. (2009). Supervised and traditional term weighting methods for automatic text categorization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(4): pp. 721–735.

[19] 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, pp. 137–142, London, UK, Springer-Verlag.

[20] Silva C. and Ribeiro B. (2003), The importance of stop word removal on recall values in text categorization, in Proceedings of the International Joint Conference on Neural Networks, 3: pp. 1661-1666, IEEE.

[21] Buckley, C. (1985). Implementation of the smart information retrieval system. Technical report, Cornell University, Ithaca, USA.

[22] Anjali, J. Ms. (2011), A Comparative Study of Stemming Algorithms, IJCTA, 2(6): pp. 1930-1938.

[23] Porter, M. F. (1980). An algorithm for suffix stripping. Program, 14(3):130–137.

[24] Mountassir, A., Benbrahim, H., and Berrada, I. (2012), An empirical study to address the problem of Unbalanced Data Sets in Sentiment Classification, In proc of IEEE International Conference on Systems, Man and Cybernetics (SMC'12), Seoul, Korea, pp. 3280-3285.

[25] Pang, B., Lee, L., and Vaithyanathain, S. (2002), Sentiment classification using machine learning techniques," In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp.79-86.

[26] Mountassir, A., Berrada, I., Benberahim, H. (2013), Representing text documents in training document spaces: a novel model for document representation, Journal of Theoretical & Applied Information Technology. 56(1), pp. 30-39. Database: Computers & Applied Sciences Complete.

[27] Dumais, S., Platt, J., Heckerman, D., and Sahami, M. (1998). Inductive learning algorithms and representations for text categorization. In Proceedings of the seventh international conference on Information and knowledge management, pp. 148–155. ACM.

[28] Scott, S. and Matwin, S. (1999). Feature engineering for text classification. In Proceedings of ICML-99, 16th International Conference on Machine Learning, pp. 379–388. Morgan Kaufmann Publishers.

[29] Nastase, V., Shirabad, J. S., and Caropreso, M. F. (2006). Using dependency relations for text classification. In Proceedings of the 19th Canadian Conference on Artificial Intelligence.

[30] Zhang, W., Yoshida, T., and Tang, X. (2008). Text classification based on multi-word with support vector machine. Knowledge-Based Systems, 21(8): pp. 879–886.

[31] Peng, X., Yi, Z., Wei, X. Y., Peng, D. Z., and Sang, Y. S. (2013). Free-gram phrase identification for modeling Chinese text. Information Processing Letters, 113(4): pp. 137–144.

[32] Zaïane, O. R. and Antonie, M. L. (2002). Classifying text documents by associating terms with text categories. In Proceedings of the 13th Australasian database conference, 5(2): pp. 215–222, Darlinghurst, Australia, Australian Computer Society, Inc.

[33] Rak, R., Stach, W., Zaïane, O. R., and Antonie, M. L. (2005). Considering re-occurring features in associative classifiers. Advances in Knowledge Discovery and Data Mining, pp. 65–72.

[34] Yang, Y. and Pedersen, J. O. (1997). A comparative study on feature selection in text categorization, Proceedings of the Fourteenth International Conference in Machine Learning, pp. 412-420.

[35] Li, Y., Hsu, D., and Chung, S. (2009). Combining multiple feature selection methods for text categorization by using rank-score characteristics, in Proceedings of 21st International Conference on Tools with Artificial Intelligence, pp. 508-517, IEEE.

[36] Shang, W., Huang, H., Zhu, H., Lin, Y., Qu, Y., & Wang, Z. (2007). A novel feature selection algorithm for text categorization. Expert Systems with Applications, 33(1): pp. 1-5.

[37] Ogura, H., Amano, H., & Kondo, M. (2009). Feature selection with a measure of deviations from Poisson in text categorization. Decision Support Systems, 36(3): pp. 6826-6832.

[38] Thomas, B. (2011). Automatic mood classification based on lyrics using various metrics, BA thesis, Faculty of Humanities, Tilburg University.

[39] Erenel, Z., Altınçay, H., and Varoğlu, E. (2011). Explicit use of term occurrence probabilities for term weighting in text categorization. Journal of Information Science and Engineering, 27(3): pp. 819–834.

[40] Altınçay, H. and Erenel, Z. (2010). Analytical evaluation of term weighting schemes for text categorization. Pattern Recognition Letters, 31: pp. 1310–1323.

[41] Ogura, H., Amano, H., and Kondo, M. (2011). Comparison of metrics for feature selection in imbalanced text classification. Expert Systems with Applications, 38(5): pp. 4978–4989.

[42] Cortes, C., and Vapnik, V. (1995). Support-vector networks. Machine Learning, 20: pp. 273-297

[43] Duda, R. O., Hart, P. E., & Stork, D.G. (2001). Pattern Classification, 2nd ed, John Wiley & Sons.

[44] Domingos, P., & Pazzani, M. (1996). Beyond independence: conditions for the optimality of the simple Bayesian classifier. Proceedings of ICML96.

[45] Yang, Y. and Liu, X. (1999). A re-examination of text categorization methods. In the proceedings of the 22nd annual international ACM SIGIR conference on Research and developments on Information Retrieval. pp 42-49.

[46] Joachims, T. (1999). Making large-scale SVM learning practical. In B. Scholkoph, C. J. C. Burges and A. J. Smola (Eds.), Advances in Kernel Methods - Support Vector Learning, Cambridge, MA: MIT Press, pp 169–184.

[47] Aizerman M. et al. (2000), Theoretical Foundations of the potential function method in pattern recognition learning, Journal of Machine Learning Research, pp. 113–141.

[48] Salton, G. (1989). Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley.

[49] Apté, C., Damerau, F., and Weiss, S. (1994). Automated learning of decision rules for text categorization. ACM Transactions on Information Systems, 12(3): pp. 233-251.

[50] Cardoso-Cachopo, A & Oliveira, A.L. (2003). An Empirical Comparison of Text Categorization. Proc. of the 10th International Symposium on String Processing and Information Retrieval, pp. 183-196.

[51] Liu, Y., Loh, H.T., & Sun, A. (2009). Imbalanced text classification: A term weighting approach. Expert Systems with Applications, 36: pp. 690-701.

[52] Debole, F. and Sebastiani, F. (2004). An analysis of the relative hardness of Reuters-21578 subsets. Journal of the American Society for Information Science and Technology, 56(6): pp. 584–596.

[53] Church, K., and Hanks, P. (1991). Word Association Norms, Mutual Information and Lexicography, Computational Linguistics, 16(1): pp. 22-29.

Appendices

Appendix A: Top 50 3-termsets obtained for the categories earn and corn of Reuters-21578 dataset.

Rank	earn	corn
1	ct, net, div	corn, maiz, sorghum
2	net, shr,div	corn, maiz, enrol
3	ct, net, record	corn, maiz, signup
4	ct, net, qtly	corn, maiz, moistur
5	ct, shr, div	corn, maiz, belt
6	ct, net, dividend	corn, maiz, thou
7	ct, shr, qtly	corn, maiz, fructos
8	shr, qtr, rev	corn, maiz, syrup
9	ct, shr, pct	corn, maiz, meal
10	ct, net, record	corn, maiz, yellow
11	ct, net, exchang	corn, maiz, fob
12	ct, shr, record	corn, maiz, harvest
13	ct, net, quarterli	corn, maiz, argentin
14	ct, shr, dividend	corn, maiz, countervail
15	rev, profit,dividend	corn, maiz, cropland
16	ct, net, payout	corn, maiz, mge
17	ct, net, payabl	corn, maiz, hfc
18	ct, net, qtly	corn, maiz, sugarcan
19	ct, shr, quarterli	corn, maiz, counselor
20	ct, shr, avg	corn, maiz, dole
21	ct, net, prior	corn, maiz, rapese
22	ct, net, qtr	corn, maiz, bale
23	ct, shr, payout	corn, maiz, cbt
24	ct, shr, prior	corn, maiz, newslett
25	ct, net, declar	corn, maiz, hrw
26	ct, shr, payabl	corn, maiz, retend
27	ct, net, shr	corn, maiz, gluten
28	ct, net, split	corn, maiz, gustafson

29	ct, qtr, div	corn, maiz, hackmann
30	ct, net, earn	corn, maiz, agrianalysi
31	ct, net, loss	corn, maiz, coars
32	rev, note, dividend	corn, maiz, susan
33	ct, net, profit	corn, maiz, srw
34	ct, net, see	corn, maiz, grasslei
35	ct, net, incom	corn, maiz, pasta
36	ct, shr, earn	corn, maiz, rudman
37	ct, shr, declar	corn, maiz, melnikov
38	ct, net, pre	corn, maiz, upheld
39	ct, qtr, dividend	corn, maiz, cst
40	ct, shr, qtr	corn, maiz, unjustifi
41	ct, net, omit	corn, maiz, tenant
42	ct, net, discontinu	corn, maiz, ae
43	ct, net, columbia	corn, maiz, graze
44	ct, net, raleigh	corn, maiz, tallow
45	ct, net, payout	corn, maiz, dn
46	ct, net, extraordinary	corn, maiz, vi
47	ct, net, ky	corn, maiz, gramm
48	ct, net, rev	corn, maiz, fieldwork
49	ct, net, auditor	corn, maiz, midmississippi
50	ct, net, restat	corn, maiz, cane

Appendix B: Top 50 3-termsets obtained for the categories alt.atheism and talk.religion.misc of 20 Newsgroups dataset.

3-termsets	alt.atheism	talk.religion.misc
1	relig, pathogen, campaign	misc, scrub, spoil
2	relig, undoubt, uneasi	misc, scrub, milton
3	exchang, fondli, pragmat	misc, scrub, finou
4	galacticent, hume, aspirin	misc, scrub, dread
5	exchang, undoubt, uneasi	misc, scrub, abhor
6	pragmat, cleric, inimit	misc, scrub, hord
7	pragmat, hume, mailer	misc, scrub, evinc
8	profit, minnestoa, mcsun	misc, scrub, fama
9	relig, undoubt, puzzl	misc, scrub, dobson
10	relig, undoubt, exodu	misc, scrub, core
11	relig, pathogen, mostli	misc, scrub, fell
12	relig, pathogen, rabbi	scrub, milton, evinc
13	relig, pathogen, mutton	scrub, milton, abhor
14	pittsburgh, motorola, willingly	scrub, milton, fell
15	pittsburgh, motorola, fabl	scrub, milton, gain
16	pittsburgh, fabl, seed	misc, psalm, chastis
17	relig, safeti, kent	misc, psalm, carrol
18	safeti, kent, sadli	misc, psalm, cohes
19	fabl, seed, evas	misc, psalm, explos
20	fabl, seed, indubit	misc, scrub, cyru
21	fabl, seed, isbn	misc, scrub, explos
22	fabl, seed, elvi	misc, scrub, elev
23	fabl, seed, extol	misc, scrub, indic
24	fabl, seed, gamma	misc, scrub, hinn
25	hospit, clarify, holi	misc, milton, india
26	seed, gamma, evas	abolit, irag, ceas
27	fabl, seed, blood	anthro, dsav, airway
28	hare, falsif, jeff	scrub, milton, psalm

29	whale, reiter, toss	misc, psalm, disord
30	reiter, toss, verg	weinss, bandwagon, anthro
31	reiter, toss, naiv	scrub, elev, weinss
32	toss, verg, seep	scrub, elev, anthro
33	gamma, evas, seep	irag, ceas, bandwagon
34	evas, seep, rabbi	coloni, erot, aver
35	evas, seep, sail	scrub, core, lewi
36	undoubt, uneasi, reiter	scrub, core, harp
37	undoubt, uneasi, toss	scrub, core, scale
38	undoubt, uneasi, verg	toronto, wall, omin
39	undoubt, uneasi, racial	grammat, interv, lynch
40	relig, uneasi, racial	misc, scrub, toronto
41	uneasi, racial, hoax	misc, scrub, mace
42	uneasi, racial, polem	misc, scrub, plant
43	racial, polem, sail	misc, scrub, bull
44	racism, serb, nott	irag, erot, aver
45	serb, nott, song	bull, aver, grand
46	nott, song, simon	erot, aver, dous
47	relig, vike, soori	misc, bull, elain
48	unsolv, rfox, stephen	hall, exercis, deja
49	vindic, rfox, stephen	heat, dealt, illumin
50	relig, nott, soori	anthro, dsav, airwai

Appendix C: Top 50 3-termsets obtained for the categories Bacterial Infections and Mycoses and Pathological Conditions, Signs and Symptoms of OHSUMED dataset.

3-termsets	Bacterial Infections and Mycoses	Pathological Conditions, Signs and Symptoms
1	bacteri, abnorm, nodal	patholog, symptom, abacteri
2	bacteri, abnorm, agent	patholog, symptom, iron
3	nodal, mycos, abdomen,	patholog, symptom, estim
4	mycos, alloxan, karnofski,	patholog, symptom, dedic
5	mycos, diseas, disciplinary	patholog, symptom, mens
6	mycos, abandon, actinomycin	patholog, symptom, lvsp
7	bacteri, mycos, abnorm	patholog, symptom, scalp
8	bacteri, mycos, lithotripsi	patholog, symptom, wast
9	mycos, abbott, gradient	patholog, symptom, west
10	abbott, gradient, infect	patholog, symptom, prdi
11	abbott, gradient, prognos	patholog, symptom, linol
12	abbott, gradient, spect	patholog, symptom, intak
13	bacteri, abnorm, spindl	patholog, symptom, nondriv
14	bacteri, abnorm, stain	patholog, symptom, wean
15	bacteri, abnorm, starv	patholog, symptom, sarn
16	bacteri, abnorm, vaccin	patholog, paracarcin, spin
17	bacteri, abnorm, zinc	symptom, paracarcin, faci
18	nodal, abdomen, oxid	symptom, paracarcin, silo
19	nodal, abdomen, pain	symptom, paracarcin, porcin
20	abnorm, gradient, pain	symptom, paracarcin, protea
21	mycos, abnorm, gradient	symptom, paracarcin, spent
22	abnorm, gradient, diseas	symptom, paracarcin, spiritu
23	abnorm, gradient, palm	symptom, paracarcin, mani
24	abnorm, gradient, meta	symptom, paracarcin, tack
25	mycos, alloxan, diseas	symptom, paracarcin, mason

26	alloxan, diseas, oxid	patient, formul, wean
27	alloxan, diseas, menier	patient, formul, imit
28	abnorm, gradient, merg	symptom, uret, acadian
29	alloxan, diseas, gamma	uret, herniotomi, acadian
30	alloxan, diseas, gravi	ethanol, doctor, cholesterol
31	merit, copi, ligat	symptom, gynaecolog, microfollicl
32	slot, tube, cereu	gynaecolog, microfollicl, extraembryon
33	garlic, bicarbon, recan	outward, patient, formul
34	yate, stapl, meta	doctor, cholesterol, fibrointim
35	bacteri, mycos, physiolog	fibrointim, imit, noct
36	bacteri, mycos, moist	nippl, petit, inlet
37	bacteri, mycos, precis	insert, pend, skin
38	bacteri, mycos, distens	sinist, pend, skin
39	bacteri, mycos, injur	pend, skin, ovin
40	bacteri, mycos, tempor	pend, skin, loco
41	bacteri, mycos, satur	corr, avct, tamoxifen
42	bacteri, mycos, abdominoperin	pressur, avct, photon
43	bacteri, mycos, limb	phakic, transluc, inson
44	bacteri, mycos, ploidi	dpti, microl, retin
45	bacteri, mycos, hygien	symptom, pend, skin
46	bacteri, mycos, zinc	symptom, corr, avct
47	bacteri, mycos, mucu	symptom, imit, noct
48	bacteri, mycos, abus	zygoma, polip, disabl
49	bacteri, mycos, chin	disastr, extraembryon, rhombic
50	bacteri, mycos, charg	inspir, flecainid, coma