Proceedings of the International Conference on Machine Vision and Machine Learning
Prague, Czech Republic, August 14-15, 2014
Paper No. 65

A Novel Scheme for Weighting Bigrams

Dima Badawi, Hakan Altınçay
Department of Computer Engineering, Eastern Mediterranean University
Famagusta, Northern Cyprus
dima.badawi@emu.edu.tr; hakan.altincay@emu.edu.tr

Abstract - In this paper, weighting bigrams in automatic binary text classification is addressed. Alternative to the conventional approach that takes into account the numbers of adjacent occurrences 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 bigrams. The document vectors are then constructed by concatenating the weight vectors of unigrams and bigrams. The proposed weighting scheme is observed to provide improved F_1 scores on two widely known benchmark domains, namely 20 Newsgroups and OHSUMED when compared to considering only the co-occurrences for assigning non-zero weights.

Keywords**:** Bigrams, Co-occurrence statistics, Term weighting, Binary text classification.

1. Introduction

With the widespread availability of online data, automatic text classification has become one of the most important tasks for effective use of a huge source of information. This problem corresponds to assigning a predefined category to a given document by taking into account its contents (Sebastiani, 2000). In order to implement this in an automated form, the documents are firstly represented as vectors where the most widely used approach is the bag-of-words (BOW). In this technique, a set of discriminative terms is firstly selected after sorting them using a selection scheme such as χ^2, Gini index or information gain (Chen et al., 2009; Liu et al., 2009; Yang et al., 2012; Bekkerman and Allan, 2004). Then, the document vectors are formed using the weights of the selected terms. In the BOW representation, the frequency of the term within the document under concern may be used as its weight. As a simpler method, binary representation may be utilized where the appearance of a term is represented as one. Experiments have shown that the product of the term frequency and a collection dependent score that is known as *collection frequency factor* generally provides better weights since the distribution of the terms in different classes is also considered. More specifically, the weight of a given term is computed as *collection frequency factor* × *collection frequency factor* of that term. In a recent study, as a collection frequency factor, relevance frequency $\left(RF = \log(2 + \frac{A}{\max(C,1)})\right)$ is proposed where A and C denote the numbers of positive and negative documents that contain the term under concern (Lan et al., 2009). It is shown to provide better scores compared to most of its competitors.

In this study, a new approach is proposed for weighting bigrams including 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 this study, we considered assigning non-zero weights to bigrams even if only one of the terms occurs. The motivation for this approach can be summarized as follows: Consider the bigram "tennis court". It can be argued that the occurrence of this bigram supports the "sports" topic. However, the occurrence of the first but not the second term is also supporting the same topic. Hence, it may be useful to assign non-zero weights to the corresponding feature in both of these cases. On the other hand, the occurrence "court" but not "tennis" may also be valuable. For instance, it may signify a different topic such as "law". More specifically, when used individually, "court"

may not be useful for differentiating between the topics "sports" and "law" since it can appear in both groups of documents. However, it becomes discriminative when evaluated together with "tennis". Based on this observation, we hypothesize that a bigram may assigned a non-zero weight even if it does not occur since it may still convey discriminative information due to partial occurrence. In this study, the co-occurrence statistics of the terms that constitute bigrams is studied to develop a better weighting scheme. The relevance frequency (*RF*) factor is updated to consider the co-occurrence statistics for generating bigram weights. Experiments conducted on two widely used benchmark datasets have shown that the proposed scheme contributes to the performance of BOW based representation. Moreover, better F_1 scores are achieved when compared to considering only the co-occurrences.

The rest of the paper is organized as follow: Section 2 reviews the related work. Section 3 discusses the details of our approach. The experimental results are presented in Section 4 and the conclusions and future work are provided in Section 5.

2. Related Work

The use of n-grams together with unigrams to achieve a better document representation has been widely addressed in the last two decades. For instance, Mladenic and Grobelnic studied the use of n-grams up to length 5 (Mladenic and Grobelnic, 1998). The experiments conducted have shown that enriched document representation with the use of n-grams together with BOW provided improved the performance for $n \leq 3$.

The number of bigrams employed by (Tan et.al., 2002) is 2% of the number of unigrams. It is shown that, with the use of a small number of bigrams, better scores can be achieved. Instead of augmenting the BOW-based representation, (Caropreso et. al., 2001) fixed the number of features to be employed and bigrams are used to replace some of the unigrams. However, they could not achieve promising results.

(Bekkerman and Allan, 2004) studied the use of bigrams together with BOW. In their study, they used mutual information for selecting discriminative bigrams. More specifically, 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 better performance scores compared to the BOW-based baseline system.

(Boulis and Ostendorf, 2005) also considered bigrams for enriched document representation and conducted their experiments on three datasets. In selecting a good set of bigrams, they quantified the additional information that each bigram brings when compared to its unigrams. They reported improved scores when compared to the BOW-based representation.

(Zhang et al., 2008) studied the use of multi-words defined as two or more consecutive terms. The main idea in this approach is to capture the context information. The multi-words are selected by comparing different sentences to find consecutive matching word sequences. The simulation experiments conducted have shown that the scores are worse compared to BOW. The use of varying lengths is also addressed by (Peng et. al., 2013). In that study, a context graph based approach is proposed to identify significant statistical phrases of arbitrary lengths. They have shown that better precision and recall scores can be achieved when compared to BOW, bigram and trigram-based representations on two different datasets.

It can be seen in the studies described above that the main problem generally addressed selection of a good subset of n-grams. In this study, we mainly focused on their weighting. The proposed scheme is presented in the following section.

3. The Proposed Weighting Scheme

Consider the binary text classification problem where there are two classes, namely positive and negative. The positive class includes the documents from the category under concern whereas the negative class includes documents from one or more other categories. Assume that t_i denotes an arbitrary unigram and $<t_i, t_j>$ denotes an arbitrary bigram.

In the case of unigram weighting, let the information elements A_1, B_1, C_1 and D_1 be defined as follows:

A_1: The number of positive documents which include t_i.
C_1: The number of negative documents which include t_i.
B_1: The number of positive documents which do not include t_i.
D_1: The number of negative documents which do not include t_i .
Then, the relevance frequency (*RF*) that is used for the computation of the collection frequency factor of a unigram is defined follows (Lan et al., 2009):

$$RF(t_i) = \log(2 + \frac{A_1}{\max(C_1,1)}) \quad (1)$$

Hence, the term weight of the unigram t_i having the frequency tf_i is computed as $tf_i \times RF(t_i)$.

In the case of bigram weighting, let the information elements A_2, B_2, C_2 and D_2 be defined as follows:

A_2: The number of positive documents which include $<t_i, t_j>$.
C_2: The number of negative documents which include $<t_i, t_j>$.
B_2: The number of positive documents which do not include $<t_i, t_j>$.
D_2: The number of negative documents which do not include $<t_i, t_j>$.
Then, RF can be updated for the computation of the collection frequency factor of a bigram as follows:

$$RF(<t_i,t_j>) = \log(2 + \frac{A_2}{\max(C_2,1)}) \quad (2)$$

In this study, we modified *RF* to take into account the occurrence of only one of the terms within the bigram for collection frequency factor computation. Let P, Q, R, S, X and Y be defined as follows:

P: The number of positive documents which include t_i but not t_j.
Q: The number of negative documents which include t_i but not t_j.
R: The number of positive documents which do not include t_i but include t_j.
S: The number of negative documents which do not include t_i but include t_j.
X: The number of positive documents which include both t_i and t_j but do not include $<t_i, t_j>$.
Y: The number of negative documents which include both t_i and t_j but do not include $<t_i, t_j>$.

It should be noted that X denotes the number of positive documents where both t_i and t_j exist but they do not appear consecutively. Then, $RF'(<t_i,t_j>)$ is defined as follows:

$$RF'(<t_i,t_j>) = \begin{cases} \log\left(2 + \frac{A_2}{\max(C_2,1)}\right), & <t_i, t_j> \text{ occurs} \\ \log\left(2 + \frac{X}{\max(Y,1)}\right), & both\ t_i \text{ and } t_j \text{ occur but } <t_i, t_j> does\ not\ occur \\ \log\left(2 + \frac{P}{\max(Q,1)}\right), & t_i \text{ occurs but } t_j \text{ does not occur} \\ \log\left(2 + \frac{R}{\max(S,1)}\right), & t_j \text{ occurs but } t_i \text{ does not occur} \\ 0, & otherwise \end{cases} \quad (3)$$

The term frequency factor is computed for each bigram as the sum of the member frequencies. Let tf_i and tf_j denote the term frequencies of the members of the 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 is computed as $(tf_i + tf_j) \times RF'(<t_i,t_j>)$.

4. Experiments

In all simulations, F_1 score is used as the performance measure. Both macro and micro F_1 scores are used to compute the overall performances within each dataset. Two widely used datasets are employed for evaluating the proposed approach, namely 20 Newsgroups and OHSUMED. 20 Newsgroups is a large corpus of 20000 newsgroup documents that are uniformly distributed among twenty different categories. It is freely available at "people.csail.mit.edu/jrennie/20Newsgroups/". OHSUMED is a subset of MEDLINE from 1987 to 1991and consists of references from 270 medical journals. The subset considered in the current study is adopted by Joachims and it includes 20000 medical abstracts (Joachims,1998). There are totally 23 categories, each corresponding to a different cardiovascular disease. Half of the corpus is used for training. For both datasets, the positive class is defined the as the target category and the negative class is defined as the set of all documents from other categories.

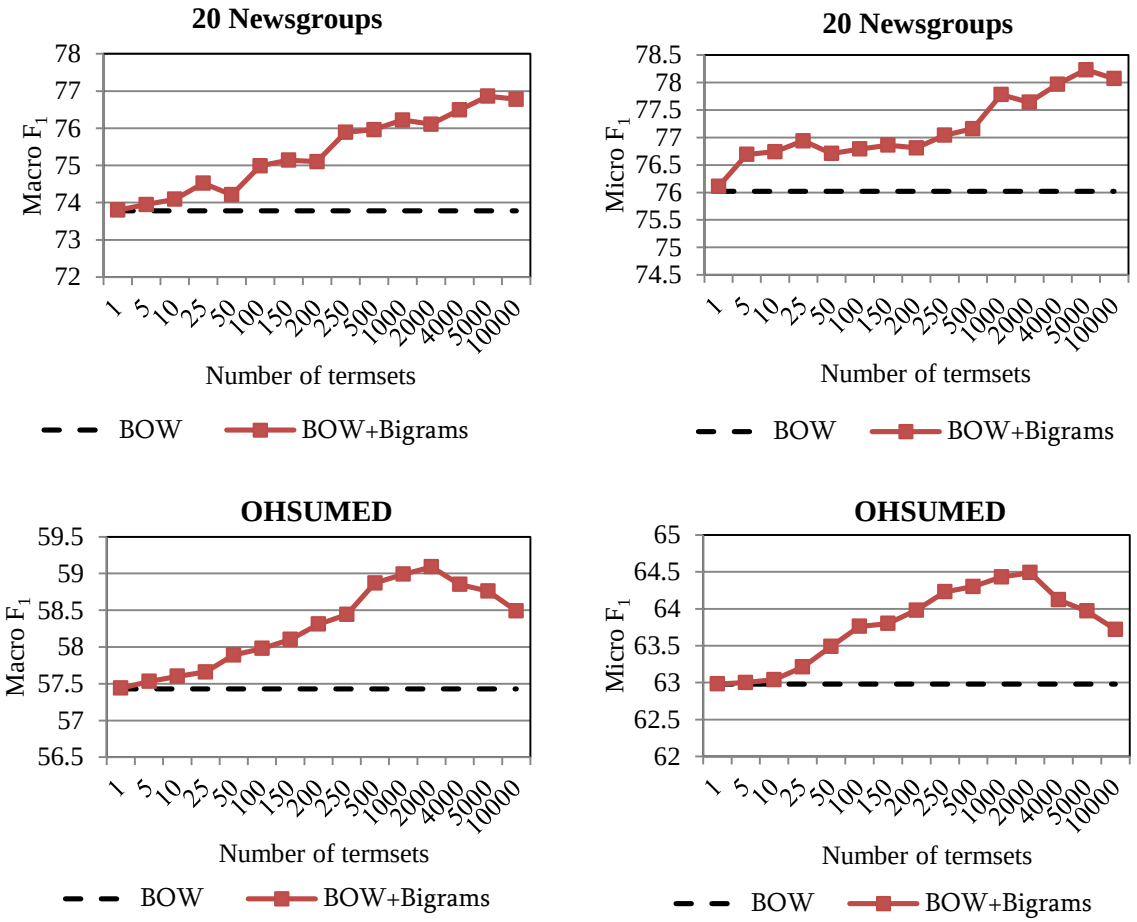

Fig. 1. The macro and micro F_1 scores achieved on 20 Newsgroups and OHSUMED datasets by using $tf_i \times RF(t_i)$ as unigram weights and $(tf_i + tf_j) \times RF(< t_i, t_j >)$ as the bigram weights.

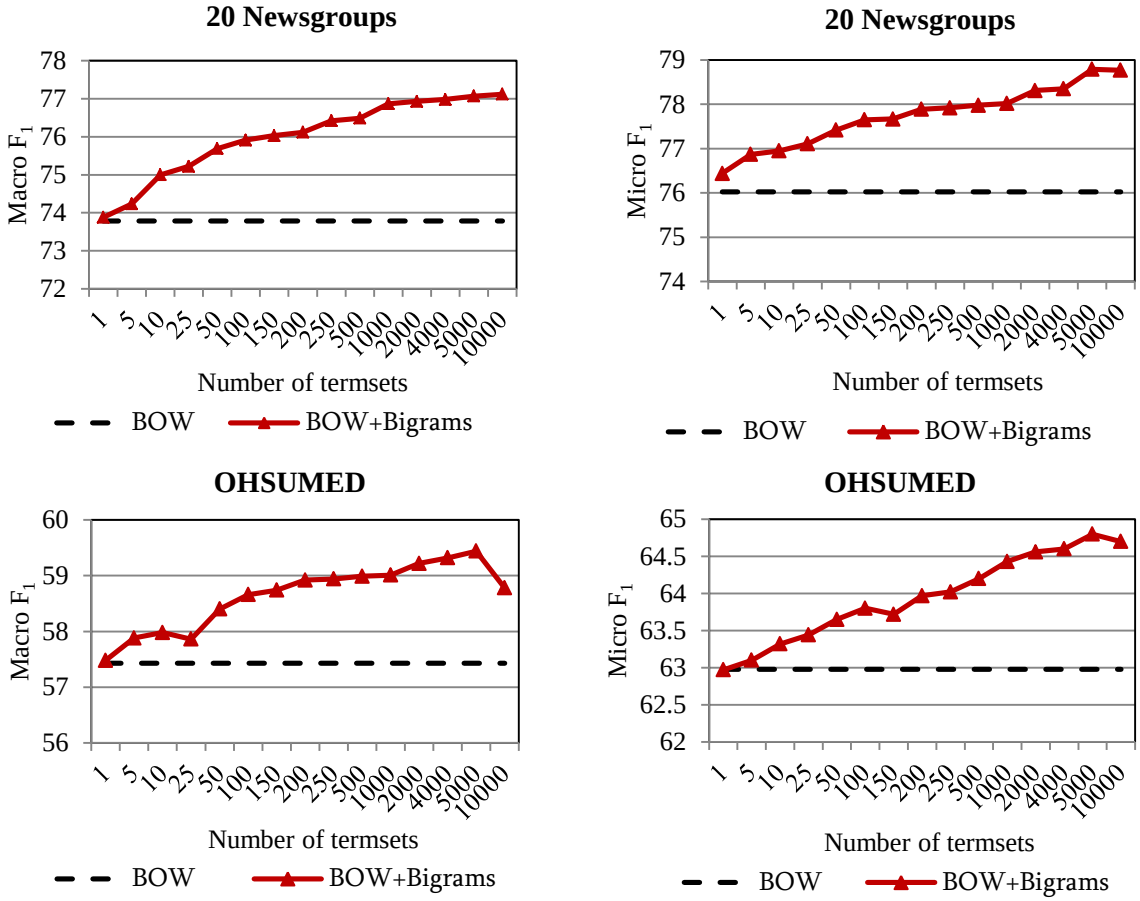

Fig. 2. The macro and micro F_1 scores achieved on 20 Newsgroups and OHSUMED datasets by using $tf_i \times RF(t_i)$ as unigram weights and $(tf_i + tf_j) \times RF'(< t_i,t_j >)$ as the bigram weights.

4.1. Experimental Setup

The Porter algorithm is firstly applied for stemming (Porter, 1980). After computing all unigrams and bigrams, SMART stoplist (Buckley, 1985) is used to eliminate stop-words 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 is eliminated from the list.

After generating the lists of unigrams and bigrams, we eliminate the bigrams that appear in less than three documents. Then, all unigrams are sorted using χ^2 defined as follows:

$$\chi^2 = \frac{N(A_1D_1 - B_1C_1)^2}{(A_1+C_1)(B_1+D_1)(A_1+B_1)(C_1+D_1)}. \quad (4)$$

The bigrams that include unigrams which are not in the top 5000 list are then discarded. The remaining bigrams are then sorted using $\chi 2$ defined as follows:

$$\chi^2 = \frac{N(A_2D_2 - B_2C_2)^2}{(A_2+C_2)(B_2+D_2)(A_2+B_2)(C_2+D_2)}. \quad (5)$$

Before computing the unigram and bigram weights, the documents lengths are normalized using cosine normalization. The normalized forms of the term frequencies are then used to compute the final forms of the weights of the unigrams and bigrams. After the document vectors are computed, the

classifier is trained using the training data. In our simulations, SVMlight toolbox with linear kernel is used for this purpose (Joachims, 1998, 1999).

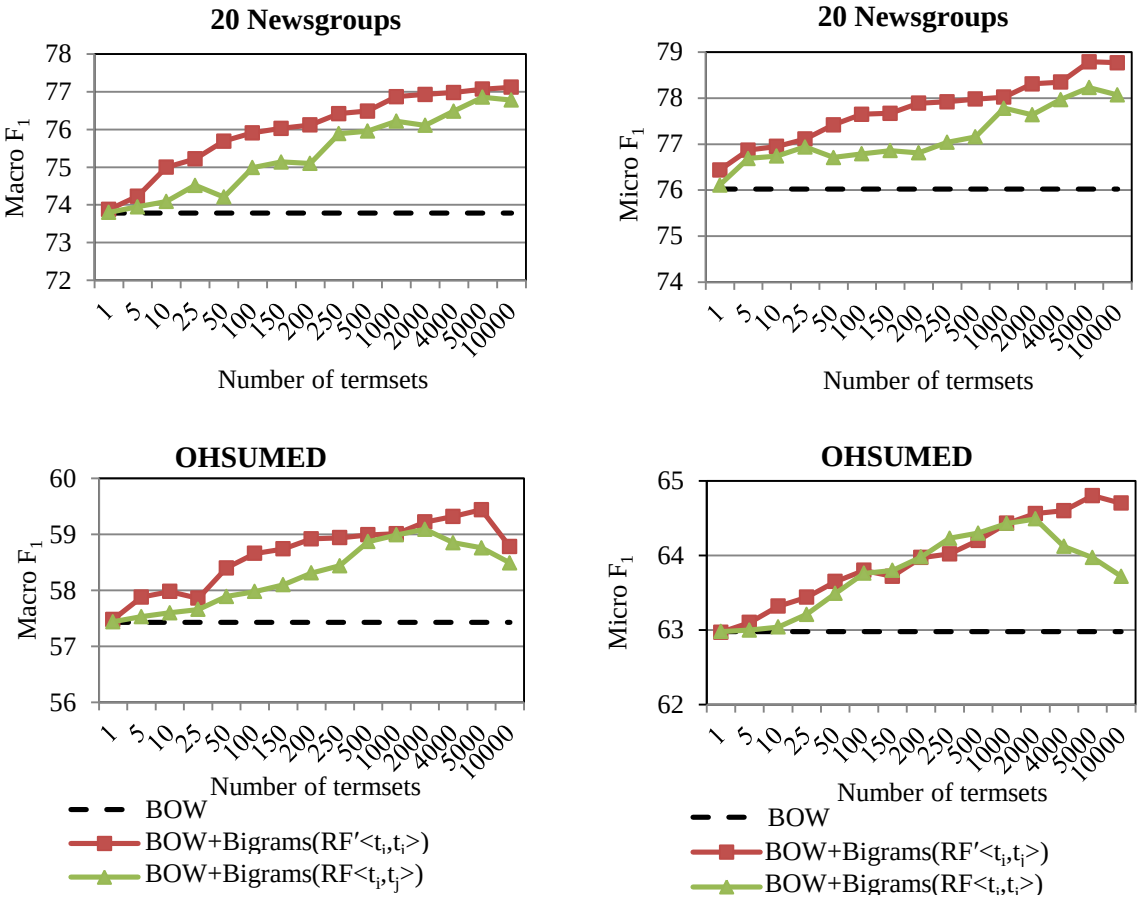

Fig. 3. The macro and micro F_1 scores achieved on 20 Newsgroups and OHSUMED by using $RF(< t_i,t_j >)$ and $RF'(< t_i,t_j >)$ as the collection frequency factors for bigrams.

4.2. Simulation Results

Fig. 1 shows the macro F_1 and micro F_1 scores obtained by using $RF(t_i)$ for unigram weighting and $RF(< t_i,t_j >)$ for bigram weighting on 20 Newsgroups and OHSUMED. The performance of the baseline BOW-based representation that employs $RF(t_i)$ for 5000 unigrams is also presented using the dashed lines for reference purposes. The horizontal axis corresponds to the number of bigrams that are concatenated with 5000 unigrams. It can be seen in the figure that the performance increases as the number of bigrams is increased up to 5000.

Fig. 2 presents the macro F_1 and micro F_1 scores achieved by using the proposed weighting scheme, $RF'(< t_i,t_j >)$ whereas the relative performances of $RF(< t_i,t_j >)$ and $RF'(< t_i,t_j >)$ are presented in Fig. 3. It can be seen that the proposed modification improves the performances on both datasets. 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 is fruitful.

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 unigrams and bigrams. The results are presented in Fig. 4. The results show that both macro and micro F_1 scores are improved on both datasets when the number of bigrams

employed is less than 500. On OHSUMED dataset, the scores drop below the baseline unigram based system when the number of bigrams is increased above 500. However, the scores achieved a far below those that are provided by the proposed scheme.

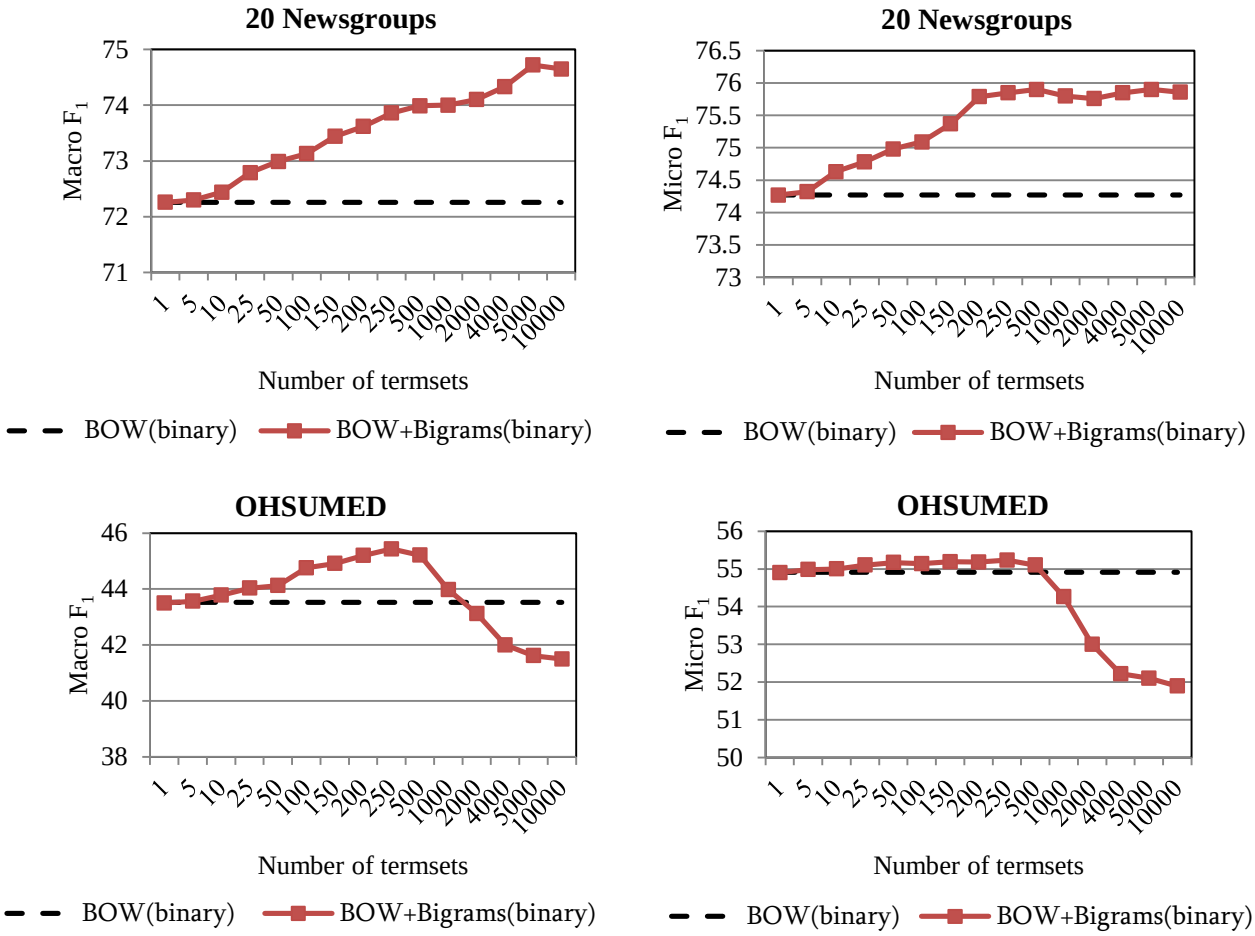

Fig. 4. The macro and micro F_1 scores achieved on 20 Newsgroups and OHSUMED using the binary representation for both unigrams and bigrams.

5. Conclusions

In this study, we present a new approach for weighting bigrams. The proposed approach is based on considering the individual occurrences of the terms within bigrams for assigning non-zero weights. Alternative to the conventional approach where the occurrences of both terms is required, the proposed scheme is based on the idea that the occurrence of only one of the terms may still convey discriminative information. By analyzing the numbers of such cases in different classes, the relevance frequency factor is modified to be employed for bigram weighting. The proposed idea is shown to provide better macro and micro F_1 scores than the conventional approach which requires both terms to appear for assigning non-zero weights.

The highest gain in macro F_1 is achieved when 5000 bigrams in addition to 5000 unigrams are employed. In particular, 4.46% ($\frac{77.07-73.78}{73.78} = 4.46$) and 3.50% improvements are achieved for 20 Newsgroups and OHSUMED respectively. On the other hand we got a gain of 2.89% for 20 Newsgroups and 2.14% if we use only 100 bigrams for both of the datasets.

In this study, we defined bigrams to be pairs of unigrams that are ranked in top 5000 using χ^2. The use of other term selection schemes and the effect of choosing a smaller set of unigrams in defining

bigrams should be further investigated. Similarly, the effectiveness of other weighting schemes for computing bigram weights using the proposed scheme should also be explored.

References

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

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, pages 9-16.

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

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, pages 78-102. IGI Publishing, Hershey, PA, USA.

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

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, pages 137-142, London, UK, UK. Springer-Verlag.

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

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):721-735.

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

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

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):137-144.

Porter, M. F. (1980). An algorithm for suffix stripping. Program, 14(3):130-137.

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

Tan, C. M., Wang, Y. F., and Lee, C. D. (2002). The use of bigrams to enhance text categorization. In Information Processing Management, volume 38, pages 529-546.

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):741-754.

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