BoWT: A Hybrid Text Representation Model for Improving Text Categorization Based on AdaBoost.MH

Bassam Al-Salemi(✉), Mohd. Juzaiddin Ab Aziz, and Shahrul Azman Mohd Noah

Knowledge Technology Research Group, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Malaysia
bassalemi@siswa.ukm.edu.my,
{juzaiddin,shahrul}@ukm.edu.my

Abstract. Text representation is the fundamental task in text categorization system. The Bag-of-Words (BoW) is a typical model for representing the texts into vectors of single words. Even though it is a simple representation model, BoW has been criticized for its disregard of the relationships between the words. Alternatively, the Latent Dirichlet Allocation (LDA) topic model has been proposed to represent the texts into a Bag-of-Topics (BoT). In LDA, the words in the corpus are statistically grouped into a small number of themes called "latent topics" in which the topics capture the semantic relationships between the words. Thus, representing the documents using BoT will dramatically accelerate the training time; as well improve the classification performance. However, BoT has been proven to not be effective for imbalanced datasets. Accordingly, this paper presents a hybrid text representation model as a combination of BoW and BoT, namely BoWT. In BoWT, the high weighted BoW's features are merged with the BoT's features to produce a new feature space. The proposed representation model BoWT is evaluated for multi-label text categorization based on the well-known boosting algorithm AdaBoost.MH. The experimental results on four benchmarks demonstrated that the BoWT representation model notably outperforms both BoW and BoT and dramatically improves the classification performance of AdaBoost.MH for text categorization.

Keywords: Text representation · Bowt · Text categorization · Adaboost.MH · Topic modeling

1 Introduction

Text representation is an essential part of any text categorization system in which the text documents are converted into a compact representation in order to be recognized by the classification algorithms. The BoW model is a standard technique of representing the documents as vectors of single words that they contain and using them as elements in the feature space. The advantage of BoW is its simplicity, as it ignores the text logical structure and layout. However, BoW has been criticized for its disregard of

C. Sombattheera et al. (Eds.): MIWAI 2016, LNAI 10053, pp. 3–11, 2016.
DOI: 10.1007/978-3-319-49397-8_1

the relationships between the words and their order among the texts. Many studies had been conducted to improve on this model to capturing the word dependency and considering the words order. Instead of considering the frequencies of the features as weights in the traditional BoW model, some weighting schemes had been proposed to tackle the features correlation problem of BoW, such as Inverse-Document-Frequency (IDF) and TFIDF [8, 12]. However, for the classification algorithms that use the binary features for inducing the classification models, e.g. AdaBoost.MH [13], the feature weighting does not make any sense.

In addition to the disregard of the words' dependencies, BoW representation model generates a vast number of features (Liu et al. 2005) and using all the extracted features for inducing the weak hypotheses of AdaBoost.MH may entail a high degree of computational time complexity, especially for large-scale datasets. That is because AdaBoost.MH produces at each boosting round a set of weak hypotheses equivalent in size to the number of the training features, refer to [4] for more details.

The high dimensionality of BoW feature space can be managed by eliminating the redundant features using an appropriate feature selection technique, such as Mutual Information, Information Gain, Chi Square-statistic, Odds Ratio, GSS Coefficient [1, 5, 6, 9–11, 14, 15]. However, feature selection may eliminate some informative features and cause information loss.

Instead of using the single words for representing the texts and training AdaBoost.MH, as BoW does, an alternative text representation model using topic modeling is proposed [3] for this task. Hence, the latent Dirichlet allocation model (LDA) [7] is used to discover the latent topics among the texts. The general outputs of LDA are; topic-word index, which contains the distribution of the words over the topics, and document-topic index, which contains the distribution of the topics over the documents. Therefore, to represent the documents into Bag-of-Topics (BoT), the document-topic index is used. This topics-based representation model has been extended to involve the most well-known multi-label boosting algorithms for multi-label text categorization [2].

Even though BoT representation model has proved to be efficient in improving text categorization based on AdaBoost.MH in general, its classification performance is poor comparing to BoW for imbalanced datasets [7]. That is because the number of topics assigned to the infrequent categories is much smaller than those assigned to the frequent categories.

Getting the advantage of feature selection for reducing the high dimensionality of BoW and selecting the high weighted features, and the advantage of BoT of capturing the semantic relationship between the words, this paper proposes a hybrid representation model as a combination of BoW and BoT. The hybrid model, which it called "**Bag of Words and Topics**" (BoWT) is proposed to tackle the limitations of both models, and to ensure increasing the number of features of the documents in the infrequent categories, as well the small texts, and give a chance to be classified correctly using AdaBoost.MH.

2 The Proposed Representation Model

The BoW is a simple model for the text representation in which the single words are used as elements to represent the texts in the feature space. However, BoW disregards the relationships between the words among the texts. Instead of using the single words in the feature space, the latent topics among the texts, which are estimated using LDA topic model, can be used. Thus, each document in the corpus is represented as a vector of topics. The advantage of using the topics as features is that the latent topic statistically clusters the words with similar meaning as one feature in the feature space. However, the BoT are not suitable for the imbalanced datasets [3]. That is because the number of topics assigned to the infrequent categories are very small in size, and that will negatively results in the classification performance. Accordingly, in this paper we proposed a hybrid representation model, namely BoWT, as a combination of BoW with BoT.

For a document d in a given corpus, d is represented using BoW as a set of words, $d = (w_1, w_2 \ldots w_n)$, and by using BoT, d is represented as a set of latent topics, $d = (t_1, t_2, \ldots t_m)$. Thus by combining both representations, d will be represented as $d = (w_1, w_2, \ldots, w_n, t_1, t_2, \ldots, t_m)$. Because the weights of both BoW and BoT are totally different; therefore, the binary weights are used for both models. As a result, the weighting of BoWT is also binary. While AdaBoost.MH uses binary features for inducing the classification model, the proposed representation model BoWT is an appropriate for this task.

To avoid the computational complexity of AdaBoost.MH training, not all the extracted features using BoW will be merged with the BoT features. Accordingly, the feature selection will be applied to reduce the size of BoW features. Thus, only the high weighted features of BoW will be combined with the latent topics in the new feature space.

3 Experiments and Results

3.1 Datasets and Experimental Settings

The datasets for multi-label text categorization which used for the evaluation purpose are: Reuters-21578 "ModApte", 20-Newsgroups (**20NG**) and **OHSUMED**. For more information about these datasets and their statistics, refer to [2]. For the Reuters-21578, the subset of 90 categories (**R90**) and the top 10 frequent categories (**R10**) are used. For each dataset the typical text preprocessing is performed: tokenization, stemming and feature selection. For feature selection, the label latent Dirichlet allocation (LLDA) is used [4]. The idea of using LLDA for feature selection is that, the features are selected based on the maximal conditional probabilities of the words across the labels, refer to [4] for more details. For LDA estimation and prediction, we followed the same settings used in [3]. However, in this paper the performance is evaluated for different numbers of topics, and the impact of using features selection before estimating the topics is also analysed. The evaluation measures used for evaluating the classification performance are: Macro-averaged F1 (MacroF1) and Micro-averaged F1 (MicroF1).

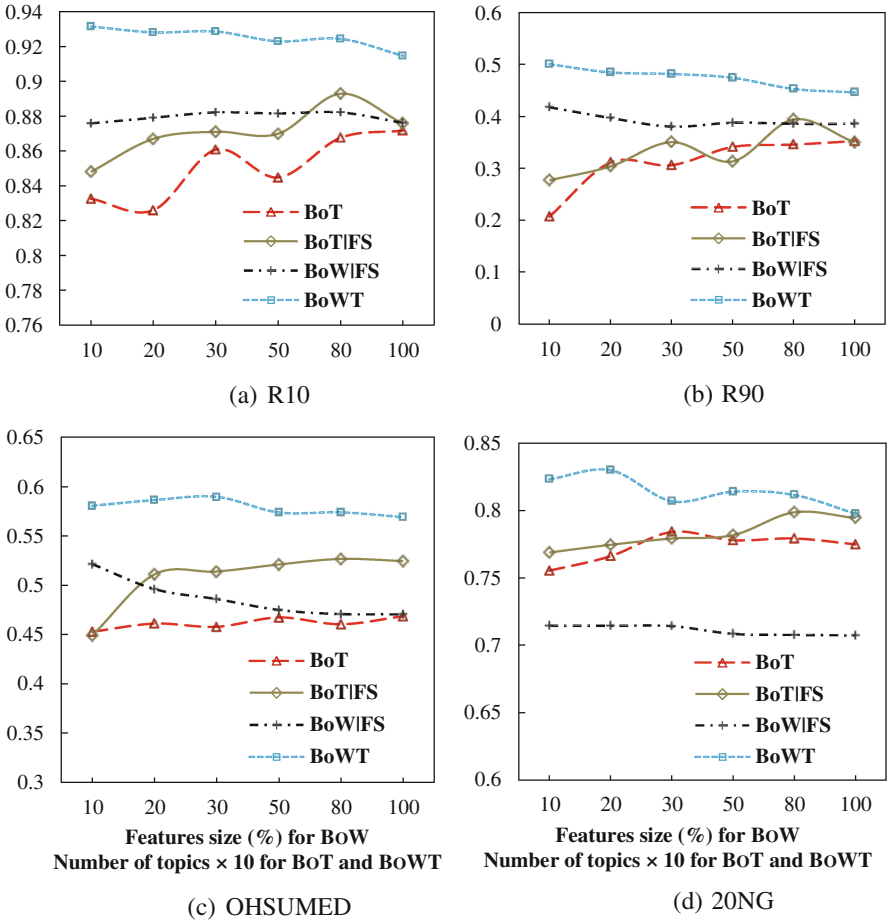

Fig. 1. The MacroF1 results of AdaBoost.MH using different text representation models

The representation models to be evaluated are:

- BoW with feature selection, dubbed (BoW|FS).
- BoT without feature selection (BoT), in which the whole extracted words from the dataset are used for LDA estimation.
- BoT after feature selection, dubbed (BoT|FS).
- BoWT, the proposed model as a combination of BoW and BoT with feature selection.

.

The BoW is evaluated on different sizes of selected features; (10, 20, 30, 50 and 80) % of the top weighted features and also 100 %, the case that all features are used without any reduction. Also BoT, BoT|FS and BoWT are evaluated with different numbers of topics: 100, 200, 300, 500, 800 and 1000 topics.

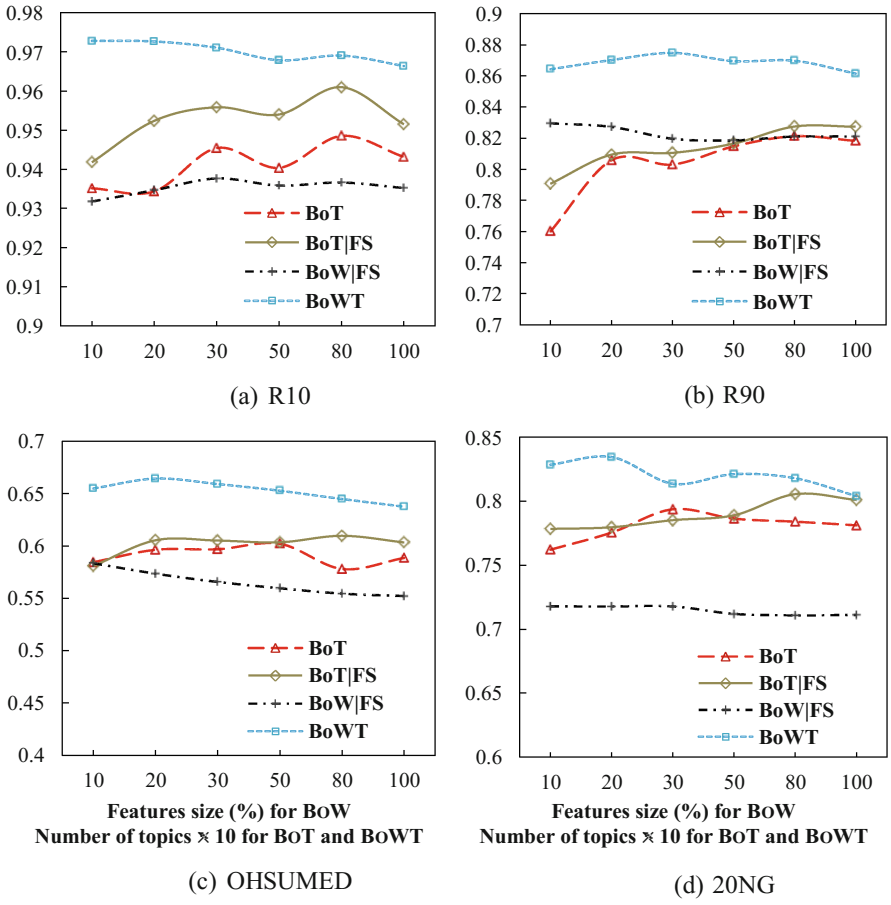

Fig. 2. The MicroF1 results of AdaBoost.MH using different text representation models

Table 1. The best MacroF1 results

Dataset	Representation	# topics	Features size	MacroF1	Rank
R10	BoT	1000	100 %	0.8717	4
	BoT|FS	800	30 %	0.8930	2
	BoWT	100	30 %	0.9315	1
	BoW|FS	–	30 %	0.8822	3
R90	BoT	1000	100 %	0.3524	4
	BoT|FS	800	10 %	0.3941	3
	BoWT	100	10 %	0.5005	1
	BoW|FS	–	10 %	0.4176	2
OHSUMED	BoT	1000	100 %	0.4686	4
	BoT|FS	800	10 %	0.5267	2

(*continued*)

Table 1. (*continued*)

Dataset	Representation	# topics	Features size	MacroF1	Rank
	BoWT	300	10 %	0.5896	1
	BoW|FS	–	10 %	0.5214	3
20NG	BoT	300	100 %	0.7841	3
	BoT|FS	800	10 %	0.7987	2
	BoWT	200	10 %	0.8299	1
	BoW|FS	–	10 %	0.7146	4

Table 2. The best MicroF1 results

Dataset	Representation	# topics	Features size	MacroF1	Rank
R10	BoT	800	100 %	0.9486	3
	BoT|FS	800	30 %	0.9610	2
	BoWT	100	30 %	0.9728	1
	BoW|FS	–	30 %	0.9377	4
R90	BoT	800	100 %	0.8212	4
	BoT|FS	800	10 %	0.8274	3
	BoWT	300	10 %	0.8747	1
	BoW|FS	–	10 %	0.8295	2
OHSUMED	BoT	500	100 %	0.6021	3
	BoT|FS	800	10 %	0.610	2
	BoWT	200	10 %	0.6644	1
	BoW|FS	–	10 %	0.5834	4
20NG	BoT	300	100 %	0.7937	3
	BoT|FS	800	10 %	0.8056	2
	BoWT	200	10 %	0.8345	1
	BoW|FS	–	10 %	0.7178	4

The classification algorithm used is the multi-label boosting algorithm AdaBoost.MH. The maximum number of iterations of AdaBoost.MH's weak learning is set to 2000 iterations.

The experiments are performed in two stages. In the first stage the BoW with feature selection and BoT are evaluated individually on all datasets. Then the best subset of BoW's features that yield the best performance is used for both BoT|FS and BoWT.

3.2 Results and Discussion

The experimental results of AdaBoost.MH classification performance measured by MacroF1 using the text representation models are illustrated in Fig. 1 for all datasets. It is clear that the proposed representation model BoWT yields the best classification

performance overall on all datasets. The BoW representation outperforms BoT|FS on average on both R10 and R90, while the best MacroF1 result using BoT|FS (0.8930) exceeds the finest MacroF1 of BoW on the R10 (0.8822). Except for R90 dataset, BoT|FS leads to the best MacroF1 overall compared to BoW. Using the BoT leads to the worst performance except for the 20NG where it exceeds BoW.

In terms of the MicroF1 results (Fig. 2), the combined representation model BoWT dramatically outperforms all other representation models on all datasets. The BoT exceeds the performance that achieved using BoW representation for all datasets except for the R90 where BoW representation obtained the best performance. Whereas, using feature selection to reduce the training features of LDA (BoT|FS) enhances the performance of topics-based representation.

The reason of the poor performance of BoT on the imbalanced dataset R90 is that the unsupervised topic model LDA takes all the documents under the training set without taking into account their categorical structure. Therefore, the documents under the infrequent categories will be represented into a few numbers of topics and that will result in AdaBoost.MH performance. The high impact of using BoT representation went to the balanced dataset, which the number of documents under each category is not varying in size, such as the 20NG. To tackle this matter both BoT and BoW representation are combined in the proposed representation BoWT. Therefore, merging the top most frequent features of BoW with the features of BoT will increase the number of informative features of the texts, particularly for the categories with small number of examples that gained small number of topics. While AdaBoost.MH uses the binary features, which the weights of the features among the texts are not considered; therefore, combining the latent topics with the word tokens will increase the classification performance.

Tables 1 and 2 summarize the best results of MacroF1 and MicroF1, respectively that obtained using different text representation models. The best MicroF1 results overall, on all datasets, are obtained when the BoWT representation model is used to represent the texts. The best MacroF1 results using BoW exceeds the results obtained using BoT on R10 and R90 datasets, while BoT leads to the best MacroF1 on OHSUMED and 20NG datasets. However, using feature selection before estimating the topics, the BoT yields the best results comparing with BoW, except for the R90 where BoW outperformed.

Regarding the best MicroF1 results (Table 2), AdaBoost.MH with the BoWT achieves the best results overall on all datasets. The BoT representation exceeds the performance of BoW on all datasets, except for the R90 where BoW yields a better performance. Moreover, reducing the features space dimension of LDA model by employing feature selection (BoT|FS), leads AdaBoost.MH to perform better than using LDA without reducing the training feature of LDA topic model (BoT).

4 Conclusion

The BoW is a typical representation model for most real-life classification problems. However, in text categorization, BoW does not capture the relationship between the words among the texts. In fact, this is the reason behind BoW simplicity. Nevertheless,

ignoring the relevance between the words may effect negatively in the classification performance, particularly for the classification algorithms that do not consider the features' weights, such as AdaBoost.MH. An alternative method to represent the text is by using the latent topics among the texts as features for inducing the classification models. Latent topics, which estimated from the text using topic modeling, are capable of capturing the semantic similarity between the words. Thus, representing the texts as a Bag-of-Topics (BoT) will improve the classification performance. However, the experimental results proved that BoT yielded a poor performance in the case of imbalanced datasets. That is because the categories with rare examples are represented into a very small number of latent topics comparing with the frequent categories. In this paper we describe a method to tackle this problem by combining the BoT's features with the high weighted features of BoW as a hybrid representation model, namely BoWT.

The experimental results demonstrate that the proposed model, BoWT, dramatically improves the classification performance of AdaBoost.MH comparing with the other models for the all datasets. The results also proved that reducing the training features of LDA topic model using feature selection increases the performance of BoT model.

References

1. Al-Salemi, B., Ab Aziz, M.J.: Statistical bayesian learning for automatic arabic text categorization. J. Comput. Sci. **7**, 39 (2010)
2. Al-Salemi, B., Ab Aziz, M.J., Noah, S.A.: Boosting algorithms with topic modeling for multi-label text categorization: a comparative empirical study. J. Inf. Sci. **41**, 732–746 (2015)
3. Al-Salemi, B., Ab Aziz, M.J., Noah, S.A.: LDA-AdaBoost.MH: Accelerated AdaBoost.MH based on latent Dirichlet allocation for text categorization. J. Inf. Sci. **41**, 27–40 (2015)
4. Al-Salemi, B., Mohd Noah, S.A., Ab Aziz, M.J.: RFBoost: an improved multi-label boosting algorithm and its application to text categorisation. Knowl.-Based Syst. **103**, 104–117 (2016)
5. Alhutaish, R., Omar, N.: Arabic text classification using k-nearest neighbour algorithm. Int. Arab J. Inf. Technol. (IAJIT) **12**, 190–195 (2015)
6. Aphinyanaphongs, Y., Fu, L.D., Li, Z., et al.: A comprehensive empirical comparison of modern supervised classification and feature selection methods for text categorization. J. Assoc. Inf. Sci. Technol. **65**, 1964–1987 (2014)
7. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
8. Dumais, S.T.: Improving the retrieval of information from external sources. Behav. Res. Methods Instrum. Comput. **23**, 229–236 (1991)
9. Duwairi, R., Al-Refai, M.N., Khasawneh, N.: Feature reduction techniques for arabic text categorization. J. Am. Soc. Inform. Sci. Technol. **60**, 2347–2352 (2009)
10. Galavotti, L., Sebastiani, F., Simi, M.: Experiments on the use of feature selection and negative evidence in automated text categorization. In: Borbinha, J., Baker, T. (eds.) ECDL 2000. LNCS, vol. 1923, pp. 59–68. Springer, Heidelberg (2000). doi:10.1007/3-540-45268-0_6

11. Lewis, D.D.: Feature selection and feature extraction for text categorization. In: Proceedings of the workshop on Speech and Natural Language. Association for Computational Linguistics, pp. 212–217 (1992)
12. Li, X., Liu, B.: Learning to classify texts using positive and unlabeled data. In: IJCAI, pp. 587–592 (2003)
13. Mukherjee, I., Schapire, R.E.: A theory of multiclass boosting. J. Mach. Learn. Res. **14**, 437–497 (2013)
14. Pekar, V., Krkoska, M., Staab, S.: Feature weighting for co-occurrence-based classification of words. In: Proceedings of the 20th International Conference on Computational Linguistics. Association for Computational Linguistics, p. 799 (2004)
15. Sebastiani, F.: Machine learning in automated text categorization. ACM Comput. Surv. (CSUR) **34**, 1–47 (2002)