

## Short Paper

---

# Using Redundancy Reduction in Summarization to Improve Text Classification by SVMs

JIAMING ZHAN AND HAN-TONG LOH

*Department of Mechanical Engineering*

*National University of Singapore*

*Singapore 119260, Singapore*

In this paper, we investigate the use of summarization technique to improve text classification. As summarization inherently assign more weights to the more important sentences in an article, this may improve the accuracy of classification of the article. Redundancy in summaries was reduced to different levels and its effect on classification performance was investigated. The classification algorithm used here was Support Vector Machines (SVMs) which has proven to be very effective and robust for text classification problem. Experimental results showed that summaries with lowest redundancy could improve the classification performance of Reuters corpus with more than 6% increase on average  $F_1$  measure. In order to explain why summarization can improve the performance while feature selection makes no sense for SVMs, a further experiment was conducted to demonstrate the difference between summarization and traditional feature selection techniques.

**Keywords:** text classification, text summarization support vector machines maximal marginal relevance, text mining

## 1. INTRODUCTION

With the explosive growth of online information, the need for us to quickly identify relevant documents is much more urgent than ever before. The technique of automatic text classification is indispensable for organizing text data. A great deal of classification algorithms have been introduced in previous researches in which Support Vector Machines (SVMs) has proved to be very effective and robust for text classification. Text summarization is an emerging technique which distills the most important information from a source to produce an abridged version for a particular user or task [8]. One technique of summarization is to assign weight to each sentence as determined by some characteristics of the sentence. Then all terms in the article are re-weighted after summarization. Terms from more important sentences will be given more weights. Some initial studies have been done to apply summarization technique in classification to re-weight terms [14, 15, 23]. However, none of prior researches considered the effect of redundant information contained in summaries which might diminish the accuracy of classification. In this paper, we reduced redundancy in the summarization process to in-

---

Received June 5, 2007; revised September 28, 2007; accepted March 20, 2008.

Communicated by Suh-Yin Lee.

investigate its effect on classification performance of SVMs. Another contribution of this paper is to clarify the difference between summarization and traditional feature selection techniques, and explain why our results showed that SVMs performance was improved by using summarization while previous studies reported that SVMs was not sensitive with feature selection [3, 12, 22].

The rest of this paper is organized as follows. Section 2 reviews previous studies on text classification, feature selection and text summarization. Section 3 describes the experimental process of using summarization with redundancy reduction in classification. Results and discussion are reported in sections 4 and 5. We also present a further experiment to show the difference between summarization and feature selection in section 5. The final section is the conclusion of the paper.

## 2. RELATED WORK

### 2.1 Text Classification and Feature Selection

Text classification, or text categorization, is to label text documents with one or more predefined categories. A lot of supervised machine learning techniques have been employed in this problem, including Naïve Bayes [18, 25], Rocchio [11], K-Nearest Neighbor [21], C4.5 [7, 9], SVMs [24]. Previous researches showed that SVMs was one of the most robust algorithms for text classification problem, outperforming other methods [5, 12, 26].

For most classifiers, including SVMs, the so-called “bag of words” representation is employed, where each document is transformed into a feature vector counting the number of occurrences of different words as features. One of the major problems of this representation scheme is the high dimensionality of the feature space (usually thousands or tens of thousands), which definitely reduces the efficiency of classification and sometimes degrades the effectiveness due to the noisy features. Feature selection is one technique to deal with such problems. Prior studies found that the accuracy for some classifiers could be improved by selecting an optimal subset from the feature space [19, 22, 27]. However, SVMs has proved to be much less sensitive to feature selection with some corpora including Reuters, a standard corpus in text categorization research. Reduction of feature space had no improvement or even small degradation on the performance of SVMs [3, 12, 22]. Joachims [12] analyzed the corpus of Reuters-21578 and found that most of the features were actually relevant for classification and feature selection might slightly reduce the accuracy.

### 2.2 Text Summarization

Summarization is the process of condensing a source text into a shorter version while preserving its information content [2]. The frameworks of most known summarization systems are similar: firstly segment the full article into several sentences (sometimes paragraphs); then assign score to each sentence according to some characteristics of this sentence; finally rank the sentences according to their scores and extract sentences with highest scores to form a summary until the expected summary length is reached [2, 17]. Sentence score can be calculated based on position, indicator phrases, word frequency, discourse structure, *etc.* or linear combination of these scores [6, 16].

The motivation of using summarization technique in text classification is this: features from more important sentences should be considered more valuable than other features and therefore should be given more weights in the “bag of words” representation. Kolcz *et al.* [15] only used summarization technique to select good features and built classifier based on the reduced feature space. The result was competitive with state-of-the-art feature selection techniques. Ko *et al.* [14] used summarization technique to calculate the importance of sentences and re-weighted all the features. Improvement was achieved on several classifiers including SVMs. Shen *et al.* [23] employed summarization technique to increase the performance of web page classification. However, none of prior researches investigated the effect on classification from redundant information in summaries. In the following experiments, we reduced redundancy in summaries to different levels in order to investigate its effect on SVMs performance. The particular method to reduce repetitive information in summary is to scale down the scores of all the sentences not yet included in the summary by an amount proportional to their similarity to the summary generated so far [20].

### 3. EXPERIMENTAL DESIGN

#### 3.1 Document Collection

The experimental document collection is based on Reuters-21578 (Distribution 1.0), which is a standard corpus widely used in text categorization research (<http://www.daviddlewis.com/resources/testcollections/reuters21578/>). The data was originally collected and labeled by the Carnegie Group, Inc. and Reuters, Ltd. There are 21578 news articles and 135 categories in this corpus. Each article may belong to one or more categories. To simplify the experiment, only the documents with single category were considered. After removing those articles with multiple labels, the remaining corpus had 9494 documents and 66 categories, in which many categories contained only one or two documents. Therefore, we selected ten most populous categories (see Table 1).

**Table 1. Ten most populous categories in Reuters-21578.**

Category	No. of articles
earn	3945
acq	2362
crude	408
trade	361
money-fx	307
interest	285
ship	158
sugar	143
coffee	116
gnp	83

After that, we removed very short articles for which summarization does not make sense, finally obtaining a corpus of 2130 documents (see Table 2). Further experiments were based on this corpus (denoted as Reuters-2130 in the rest of the paper).

**Table 2. The corpus Reuters-2130 used in our experiments.**

Category	No. of articles
acq	670
earn	499
trade	255
crude	234
money-fx	125
interest	82
ship	74
sugar	72
coffee	67
gnp	52

### 3.2 Summarization Process

There are two kinds of summaries: extracts and abstracts. An extract is a selection of some of the material of the original, while an abstract is a condensation and reformulation of the original document [10]. In this experiment, we employed extraction rather than abstraction due to the characteristics of simplicity, robustness and domain independency of extraction.

Our purpose is to investigate the redundant information in summarization and its effect on text classification performance. The method of Maximal Marginal Relevance (MMR) was used to reduce the redundancy in summaries. The initial goal of MMR was to reduce redundancy while preserving query relevance in re-ordering retrieved articles in information retrieval system [4]. Similar to information retrieval process which ranks documents by relevance to query, summarization can be viewed as a process to rank sentences by relevance to the overall document. Therefore, MMR can also be used in summarization to reduce redundancy. In this experiment, we used MMR-based summarization method to generate summaries. The summarization process was as follows:

- Pick up the first sentence of the article and add it into summary  $S$ . For news articles in Reuters corpus, the first sentence usually carries very important topic information about the whole article.
- Calculate Marginal Relevance for each sentence in  $D-S$  ( $D$  is the whole document,  $D-S$  is the set of sentences in  $D$  which have not been included into  $S$ ). The definition of Marginal Relevance for sentence  $s_i$  is as follows:

$$MR(s_i) = \lambda Sim(s_i, D) - (1 - \lambda) \max_{s_j \in S} Sim(s_i, s_j), \text{ where } 0 \leq \lambda \leq 1.$$

The first part  $Sim(s_i, D)$  is the similarity between  $s_i$  and the whole article  $D$ , which indicates the relevance of this sentence to the main topic of this article. In the second part,  $Sim(s_i, s_j)$  is the similarity between  $s_i$  and  $s_j$  ( $s_j$  is any sentence from summary  $S$ ). Therefore, the second part measures the repetitive information carried by sentence  $s_i$  with respect to the summary  $S$ . The higher this value, more redundant information is contained in this sentence. With regard to  $Sim$ , we adopt a cosine similarity measure between sen-

tence vectors. Each element of a sentence vector represents the weight of a word-stem in the sentence after removing stop words.

- Pick up the sentence with maximal value of Marginal Relevance and add it into  $S$ .
- Repeat the 2nd and 3rd steps until expected summary length is reached. In this experiment, summary length is set to 10% of the original article, *i.e.* compression ratio is 10%. Under this compression ratio, summaries for most articles in Reuters-2130 contain 2 or more sentences. Initial experiments showed that summaries with compression ratio of 10% achieved better performance in classification than those with higher compression ratios of 25%, 50%, 75%, and very short summaries like one sentence or title. Therefore, compression ratio was fixed at 10% and we tuned  $\lambda$  value to study the effect of redundancy on classification.

From the above process we can see, the redundancy contained in the final summary is tuned by  $\lambda$  value, which is ranged from 0 to 1. When  $\lambda$  is 1, the second part of Marginal Relevance equals 0 and this process is similar to traditional summarization without redundancy reduction. When  $\lambda$  decreases to 0, the first part of Marginal Relevance is 0 and redundancy in the final summary is reduced to minimal. Therefore,  $\lambda$  value actually indicates the level of redundancy contained in the final summary. In this experiment, we generated summaries for all articles from corpus Reuters-2130 based on  $\lambda$  values of 0, 0.3, 0.7, 1 and used these summaries for further classification tasks. Two examples of summaries based on  $\lambda = 0$  and  $\lambda = 1$  are listed here:

Summaries  $\lambda = 0$ :

Dome Petroleum Ltd's proposal to restructure debt of more than 6.10 billion Canadian dlrs includes provisions that may force the company to sell its 42 pct stake in <Encor Energy Corp Inc>, Dome said in a U.S. Securities and Exchange Commission filing. "However, the final outcome of the negotiations cannot be predicted at this time," it said.

Summaries  $\lambda = 1$ :

Dome Petroleum Ltd's proposal to restructure debt of more than 6.10 billion Canadian dlrs includes provisions that may force the company to sell its 42 pct stake in <Encor Energy Corp Inc>, Dome said in a U.S. Securities and Exchange Commission filing. Dome said in the filing that its debt plan proposes making payments under a five year income debenture to the lender whose debt is secured by Dome's Encor shares.

### 3.3 SVMs Classification

For classification, we chose SVMs since it has proved to be more suitable and robust than other classifiers for text categorization problem [12, 26]. Based on the Structural Risk Minimization principle, SVMs can find a hyperplane  $\vec{w}^T \vec{x} + b = 0$  that "best" separates data points from two classes by using only a few of training points, known as support vectors. The detailed discussion of SVMs can be found in the book of [24]. The simple linear kernel SVMs was used in our experiment because previous studies showed that it was very effective for Reuters data set [5]. The tool used here was SVM<sup>light</sup> [13].

In order to reduce the uncertainty of data partition for training and testing set, 5-fold cross validation procedure was applied in our experiment. The whole data set of Reuters-

2130 was randomly split into 5 mutually exclusive subsets or “folds”. Training and testing process were performed 5 times. In each iteration, 4 folds out of 5 were combined as a training set and the last fold was used as testing set. Some preprocessing such as stop words removal and term stemming were performed before classification.

To evaluate the classification performance, we used macro  $F_1$  measure. For each category, a confusion matrix after classification is given in Table 3.

**Table 3. Confusion matrix for classification.**

		Actual	
		Positive	Negative
Predicted	Positive	$a$	$b$
	Negative	$c$	$d$

The definition of accuracy (AC), precision (P), recall (R),  $F_1$  measure are:

$$AC = \frac{a+d}{a+b+c+d}, P = \frac{a}{a+b}, R = \frac{a}{a+c}, F_1 = \frac{2PR}{P+R}.$$

Macro  $F_1$  measure is the average  $F_1$  value over all the categories.

#### 4. EXPERIMENTAL RESULTS

Table 4 shows classification result of iteration 1 in 5-fold cross validation. Classification results based on original full articles are recorded in column “Original”. The last four columns record the classification results based on summaries when  $\lambda$  value ranges from 0 to 1. Average  $F_1$  value over all categories (macro  $F_1$  measure) are given at the end of this table.

**Table 4. SVMs classification results for iteration 1 in 5-fold cross validation (percent).**

Category	Original	Summaries			
		$\lambda = 0$	$\lambda = 0.3$	$\lambda = 0.7$	$\lambda = 1$
acq	90.62871	93.24273	93.7482	93.24273	93.77757
coffee	88.37417	97.8706	97.8706	97.8706	95.65399
crude	87.76979	89.04915	89.20815	89.35764	89.20815
earn	84.21118	87.12872	83.78596	86.37644	81.9122
gnp	75.00292	78.26405	69.56344	63.63636	69.56344
interest	60.46792	61.90706	61.90706	54.99891	52.38085
money-fx	62.06894	71.87527	58.17912	62.06894	55.55598
ship	0	11.10691	11.10691	5.712065	11.10691
sugar	85.71429	97.8706	97.8706	97.8706	97.8706
trade	87.49979	82.63152	84.8839	84.02598	84.2136
Macro $F_1$	72.17377	77.09466	74.8124	73.51603	73.12433

From this table we can see that classification performance for some categories was greatly improved through summarization, such as categories “coffee” and “sugar”. Minor

improvement or degradation was obtained for other categories. Summaries with  $\lambda = 0$  offered the best classification performance, which achieved better performance than full articles for all categories except the category “trade”. On average, summaries with  $\lambda = 0$  achieved 77.1% of  $F_1$  value and about 5% more than full articles, and was also better than the performance of other  $\lambda$  values.

Classification results for all iterations in 5-fold cross validation are presented in Table 5 and Fig. 1. The average  $F_1$  values for all iterations are given at the end of Table 5. From the results we found that for all iterations, summaries made better classification performance than full articles and the best accuracy was achieved when  $\lambda = 0$ , *i.e.* redundancy in summaries was reduced to minimal. On average, summaries with  $\lambda = 0$  could improve SVMs performance with more than 6% increase on  $F_1$  measure. The results also showed that lower  $\lambda$  value tended to generate higher classification performance, *i.e.* redundancy reduction in summaries was helpful for text classification.

**Table 5. SVMs classification results for all iterations (percent).**

Iterations	Original	Summaries			
		$\lambda = 0$	$\lambda = 0.3$	$\lambda = 0.7$	$\lambda = 1$
1	72.17377	77.09466	74.8124	73.51603	73.12433
2	70.81778	76.96226	75.02443	75.6348	75.10113
3	66.3942	74.27021	74.19104	73.57256	72.18078
4	69.83592	79.0196	77.91898	77.14433	75.09977
5	73.67756	78.04514	76.10518	76.74656	74.53137
Average	70.57985	77.07837	75.61041	75.32286	74.00748

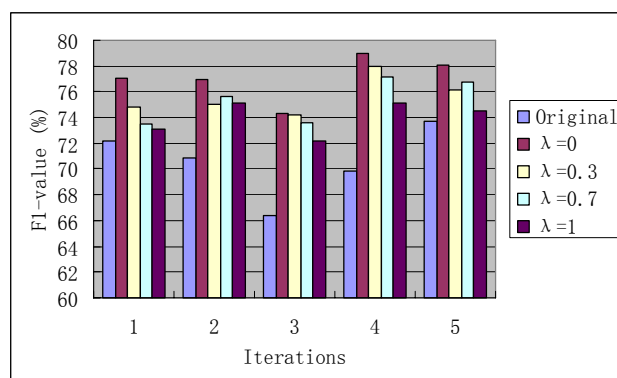


Fig. 1. SVMs classification results for all iterations.

## 5. FURTHER DISCUSSION

From above results, summarization could improve the classification performance and redundancy reduction in summaries was helpful for SVMs algorithm. These results seem to contradict with previous studies which reported that feature selection was not effective with SVMs and might even degrade the classification performance on Reuters corpus (Joachims 1998). In fact, summarization process selects an optimal feature set

and re-weights all the features. This process is different from traditional feature selection techniques which only rank all the features according to their importance and select top ones. The following experiment is to clarify this problem.

After stop words removal and term stemming, Reuters-2130 has 7889 features (noted as FS-7889). After summarization with  $\lambda = 0$  and compression ratio of 10%, the dimension of feature space reduces to 3871 (noted as FS-3871). The difference between FS-7889 and FS-3871 are 4018 features, which is noted as FS-4018. We conducted classification of original full articles based on FS-7889, FS-3871 and FS-4018 (The training and testing set were the same with iteration 1 in the previous 5-fold cross validation procedure). The results are presented in Table 6. From the results we found that classifier trained on FS-3871 achieved 71.7% of macro  $F_1$  score, which is slightly lower than that of FS-7889 (72.2%) and much better than that of FS-4018, which is only 37.2%. This is probably because summarization actually selected an optimal subset (FS-3871) of features from the whole corpus (FS-7889) and FS-4018 contains most of the noisy features. The results were consistent with previous researches which showed that SVMs was not sensitive and even had minor degraded performance with feature selection [12].

**Table 6. Comparison between summarization and feature selection.**

Category	Original	Original	Original	Summaries
acq	90.62871	90.13607	73.48134	93.24273
coffee	88.37417	90.90711	28.57633	97.8706
crude	87.76979	87.76979	62.99055	89.04915
earn	84.21118	85.62421	60.00054	87.12872
gnp	75.00292	75.00292	0	78.26405
interest	60.46792	49.99724	0	61.90706
money-fx	62.06894	56.60325	40.00284	71.87527
ship	0	5.712065	5.712065	11.10691
sugar	85.71429	85.71429	49.99625	97.8706
trade	87.49979	89.38258	51.12552	82.63152
Macro $F_1$	72.17377	71.68495	37.18854	77.09466

We also list the classification result of summaries ( $\lambda = 0$  and 10% compression ratio) based on feature space FS-3871. We found that based on the same feature space of FS-3871, summaries achieved better classification performance than original full articles. The reason is that although based on the same feature space, summaries and full articles will have different weights for each feature. The results showed that summaries offered a better weighting scheme for text classification.

To sum up, summarization is actually a process with the combination of feature selection and feature re-weighting. Through this process, better SVMs classification performance is achieved on Reuters-2130.

## 6. CONCLUSION

In this paper, we applied summarization technique to text classification on a subset of Reuters-21578 collection. We reduced the redundant information contained in the

summaries to investigate its effect on classification performance. Experimental results showed that redundancy reduction was helpful for classification and summaries with lowest redundancy could improve the classification performance of Reuters collection with more than 6% on average  $F_1$  measure.

We trained classifiers using original articles based on original feature space FS-7889 and reduced feature space FS-3871. The results showed that FS-3871 generated a slightly lower  $F_1$  score than FS-7889. This was consistent with previous studies which reported that SVMs was not sensitive with feature selection with respect to Reuters corpus. We also found that classifier trained using summaries is much better than classifier trained using original full articles based on the same feature space FS-3871, which means that summarization can actually re-weight all the features and this re-weighting process is helpful for SVMs classification.

## REFERENCES

1. C. Apte, F. Damerau, and S. M. Weiss, "Automated learning of decision rules for text categorization," *ACM Transactions on Information Systems*, Vol. 12, 1994, pp. 233-251.
2. R. Barzilay and M. Elhadad, "Using lexical chains for text summarization," in *Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, 1997, pp. 10-17.
3. J. Brank, *et al.*, "Interaction of feature selection methods and linear classification models," in *Proceedings of the International Conference on Machine Learning, Workshop on Text Learning*, 2002, pp. 505-509.
4. J. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking for re-ordering documents and producing summaries," in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1998, pp. 335-336
5. S. Dumais, *et al.*, "Inductive learning algorithms and representations for text categorization," in *Proceedings of the 7th International Conference on Information and Knowledge Management*, 1998, pp. 148-155.
6. H. P. Edmundson, "New methods in automatic extracting," *Journal of the Association for Computing Machinery* Vol. 16, 1969, pp. 264-285.
7. E. Gabrilovich and S. Markovitch "Text categorization with many redundant features: using aggressive feature selection to make SVMs competitive with C4.5," in *Proceedings of the 21st International Conference on Machine Learning*, 2004, pp. 321- 328.
8. Y. Guo and G. Stylios, "An intelligent summarization system based on cognitive psychology," *Information Sciences*, Vol. 174, 2005, pp. 1-36.
9. T. Hidekazu, *et al.*, "Estimating sentence types in computer related new product bulletins using a decision tree," *Information Sciences*, Vol. 168, 2004, pp. 185-200.
10. E. Hovy and C. Y. Lin, "Automated text summarization in SUMMARIST," *Advances in Automatic Text Summarization*, 1999, pp. 81-94.
11. T. Joachims, "A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization," in *Proceedings of the 14th International Conference on Machine*

- Learning*, 1997, pp. 143-151.
12. T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Proceedings of the 10th European Conference on Machine Learning*, 1998, pp. 137-142.
  13. T. Joachims, "Making large-scale support vector machine learning practical," *Advances in Kernel Methods: Support Vector Machines*, 1999, pp. 169-184.
  14. Y. Ko, *et al.*, "Automatic text categorization using the importance of sentences," in *Proceedings of the 19th International Conference on Computational Linguistics*, Vol. 1, 2002, pp. 1-7.
  15. A. Kolcz, *et al.*, "Summarization as feature selection for text categorization," in *Proceedings of the 10th International Conference on Information and Knowledge Management*, 2001, pp. 365-370.
  16. H. P. Luhn, "The automatic creation of literature abstracts," *IBM Journal of Research and Development*, Vol. 2, 1958, pp. 159-165.
  17. D. Marcu, "Discourse trees are good indicators of importance in text," *Advances in Automatic Text Summarization*, 1999, pp. 123-136.
  18. A. McCallum and K. Nigam, "A comparison of event models for Naïve Bayes text classification," in *Proceedings of AAAI Workshop on Learning for Text Categorization*, 1998, pp. 41-48.
  19. D. Mladenic and M. Grobelnik, "Feature selection for unbalanced class distribution and Naïve bayes," in *Proceedings of the 16th International Conference on Machine Learning*, 1999, pp. 258-267.
  20. D. R. Radev, *et al.*, "Experiments in single and multi-document summarization using MEAD," in *Proceedings of Document Understanding Conference Workshop on Text Summarization*, 2001, pp. 112-117.
  21. I. Rahal and W. Perrizo, "An optimized approach for KNN text categorization using P-trees," in *Proceedings of ACM Symposium on Applied Computing*, 2004, pp. 613-617.
  22. M. Rogati and Y. Yang, "High-performing feature selection for text classification," in *Proceedings of the 11th International Conference on Information and Knowledge Management*, 2002, pp. 659-661.
  23. D. Shen, *et al.*, "Web-page classification through summarization," in *Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieval*, 2004, pp. 242-249.
  24. V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, 1995.
  25. R. R. Yager, "An extension of the naïve Bayesian classifier," *Information Sciences*, Vol. 176, 2006, pp. 577-588.
  26. Y. Yang and X. Liu, "A re-examination of text categorization methods," in *Proceedings of Special Interest Group on Information Retrieval*, 1999, pp. 42-49.
  27. Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *Proceedings of the 14th International Conference on Machine Learning*, 1997, pp. 412-420.

**Jiaming Zhan (詹家明)** He is currently a staff Engineer in IBM Singapore and a Ph.D. candidate in the Department of Mechanical Engineering at the National University of Singapore. His research interests include data mining, information and knowledge management. He received his Bachelor in Mechanical Engineering and Bachelor in Information Science from the University of Science and Technology of China.

**Han-Tong Loh** received his Bachelor in Engineering from the University of Adelaide, his Master of Engineering from the National University of Singapore (NUS) and his Master of Science and Ph.D. in Mechanical Engineering from the University of Michigan. He is an associate professor and Deputy Head in the Department of Mechanical Engineering at NUS. He is also a Fellow of the Singapore-MIT Alliance, which is an innovative engineering education and research collaboration between MIT, NUS and the Nanyang Technological University, to promote global education and research in engineering. His research interests include data mining, rapid prototyping, robust design and computer aided design.