

# Co-training for Semi-supervised Sentiment Classification Based on Dual-view Bags-of-words Representation

Rui Xia<sup>1,2</sup>, Cheng Wang<sup>1</sup>, Xinyu Dai<sup>2</sup>, and Tao Li<sup>3,4</sup>

<sup>1</sup>School of Computer Science, Nanjing University of Science & Technology, China

<sup>2</sup>State Key Laboratory for Novel Software Technology, Nanjing University, China

<sup>3</sup>School of Computer Science, Florida International University, USA

<sup>4</sup>School of Computer Science, Nanjing University of Posts & Telecommunications, China

rxia@njust.edu.cn, wangcheng1022@gmail.com,

daixinyu@nju.edu.cn, taoli@cs.fiu.edu

## Abstract

A review text is normally represented as a bag-of-words (BOW) in sentiment classification. Such a simplified BOW model has fundamental deficiencies in modeling some complex linguistic phenomena such as negation. In this work, we propose a dual-view co-training algorithm based on dual-view BOW representation for semi-supervised sentiment classification. In dual-view BOW, we automatically construct antonymous reviews and model a review text by a pair of bags-of-words with opposite views. We make use of the original and antonymous views in pairs, in the training, bootstrapping and testing process, all based on a joint observation of two views. The experimental results demonstrate the advantages of our approach, in meeting the two co-training requirements, addressing the negation problem, and enhancing the semi-supervised sentiment classification efficiency.

## 1 Introduction

In the past decade, there has been an explosion of user-generated subjective texts on the Internet in forms of online reviews, blogs and microblogs. With the need of automatically identifying sentiments and opinions from those online texts, sentiment classification has attracted much attention in the field of natural language processing.

Lots of previous research focused on the task of supervised sentiment classification. However, in some domains, it is hard to obtain a sufficient amount of labeled training data. Manual annotation is also very expensive and time-consuming. To address this problem, semi-supervised learning

approaches were employed in sentiment classification, to reduce the need for labeled reviews by taking advantage of unlabeled reviews.

The dominating text representation method in both supervised and semi-supervised sentiment classification is known as the bag-of-words (BOW) model, which is difficult to meet the requirements for understanding the review text and dealing with complex linguistic structures such as negation. For example, the BOW representations of two opposite reviews “It works well” and “It doesn’t work well” are considered to be very similar by most statistical learning algorithms.

In supervised sentiment classification, many approaches have been proposed in addressing the negation problem (Pang et al., 2002; Na et al., 2004; Polanyi and Zaenen, 2004; Kennedy and Inkpen, 2006; Ikeda et al., 2008; Li et al., 2010b; Orimaye et al., 2012; Xia et al., 2013). Nevertheless, in semi-supervised sentiment classification, most of the current approaches directly apply standard semi-supervised learning algorithms, without paying attention to appropriate representation for review texts. For example, Aue and Gamon (2005) applied the naïve Bayes EM algorithm (Nigam et al., 2000). Goldberg and Zhu (2006) applied a graph-based semi-supervised learning algorithm by (Zhu et al., 2003). Wan (2009) employed a co-training approach for cross-language sentiment classification. Li et al. (2010a) employed co-training with personal and impersonal views. Ren et al. (2011) explored the use of label propagation (Zhu and Ghahramani, 2002).

As pointed by (Goldberg and Zhu, 2006): it is necessary to investigate better review text representations and similarity measures based on linguistic knowledge, as well as reviews’ sentiment patterns. However, to the best knowledge, such investigations are very scarce in the research of semi-

supervised sentiment classification.

In (Xia et al., 2013), we have developed a dual sentiment analysis approach, which creates antonymous reviews and makes use of original and antonymous reviews together for supervised sentiment classification. In this work, we propose a dual-view co-training approach based on dual-view BOW representation for semi-supervised sentiment classification. Specifically, we model both the original and antonymous reviews by a pair of bags-of-words with opposite views. Based on such a dual-view representation, we design a dual-view co-training approach. The training, bootstrapping and testing processes are all performed by observing two opposite sides of one review. That is, we consider not only how positive/negative the original review is, but also how negative/positive the antonymous review is.

In comparison with traditional methods, our dual-view co-training approach has the following advantages:

- Effectively address the negation problem;
- Automatically learn the associations among antonyms;
- Better meet the two co-training requirements in (Blum and Mitchell, 1998).

## 2 Related Work

The mainstream of the research in sentiment classification focused on supervised and unsupervised learning tasks. In comparison, semi-supervised sentiment classification has much less related studies. In this section, we focus on reviewing the work of semi-supervised sentiment classification.

Aue and Gamon (2005) combined a small amount of labeled data with a large amount of unlabeled data in target domain for cross-domain sentiment classification based on the EM algorithm. Goldberg and Zhu (2006) presented a graph-based semi-supervised learning algorithm (Zhu et al., 2003) for the sentiment analysis task of rating inference. Dasgupta and Ng (2009) proposed a semi-supervised approach to mine the unambiguous reviews at first and then exploiting them to classify the ambiguous reviews, via a combination of active learning, transductive learning and ensemble learning. Ren et al. (2011) explored the use of label propagation (LP) (Zhu and Ghahramani, 2002) in building a semi-supervised sentiment classifier, and compared their results with Transductive SVMs(T-SVM). LP and T-SVM are

transductive learning methods where the test data should participate in the training process.

Zhou et al. (2010) proposed a deep learning approach called active deep networks to address semi-supervised sentiment classification with active learning. Socher et al. (2012) introduced a deep learning framework called semi-supervised recursive autoencoders for predicting sentence-level sentiment distributions. The limitation of deep learning approaches might be their dependence on a considerable amount of unlabeled data to learn the representations and the inability to explicitly model the negation problem.

One line of semi-supervised learning research is to bootstrap class labels using techniques like self-training, co-training and their variations. Wan (2009) proposed a co-training approach to address the cross-lingual sentiment classification problem. They made use of the machine translation service to produce two views (a English view and a Chinese view) for co-training a Chinese review sentiment classifier, based on English corpus and unlabeled Chinese corpus. Li et al. (2010a) proposed an unsupervised method at first to automatically separate the review text into a personal view and an impersonal view, based on which the standard co-training algorithm is then applied to build a semi-supervised sentiment classifier. Li et al. (2011) further studied semi-supervised learning for imbalanced sentiment classification by using a dynamic co-training approach. Su et al. (2012) proposed a multi-view learning approach to semi-supervised sentiment classification with both feature partition and language translation strategies (Wan, 2009). Following (Li et al., 2010a), Li (2013) proposed a co-training approach which exploits subjective and objective views for semi-supervised sentiment classification. Our approach can also be viewed as a variation of co-training. The innovation of our approach is the dual-view construction technique by incorporating antonymous reviews and the bootstrapping mechanism by observing two opposite sides of one review.

## 3 The Proposed Approach

### 3.1 Dual-view BOW Representation for Review Texts

Every coin has two sizes. In this work, we are motivated to automatically construct the antonymous reviews, consider the original and antonymous reviews as two opposite sides of one review, and rep-

Feature Space	Original View	Anonymous View
app	1	1
phone	1	1
didn't	1	0
work	1	1
well	1	1
disappointing	1	0
recommend	1	1
satisfactory	0	1

Figure 1: An illustration of the dual-view BOW representation. The feature vector with black font color and grey background denotes the original view; while the one with white font color and black background denotes the reversed anonymous view.

resent them in pairs by a dual-view BOW model.

Look at the following example:

**Original Review:** “The app doesn’t work well on my phone. Disappointing. Don’t recommend it.”

**Anonymous Review:** “The app works well on my phone. Satisfactory. Recommend it.”

Given an original review, its anonymous review is automatically created as follows<sup>1</sup>: 1) We first detect the negations in each subsentence of the review text; 2) If there is a negation, we remove negators in that subsentence; 3) Otherwise, we reverse all the sentiment words in the subsentence into their antonyms, according to a pre-defined antonym dictionary<sup>2</sup>.

We subsequently use a dual-view BOW model to represent such a pair of reviews, as shown in Figure 1. The original and anonymous reviews will be used in pairs in our dual-view semi-supervised learning approach. As we determine the sentiment of one review, we could observe not only the original view, but also the anonymous view.

<sup>1</sup>It is worth noting that our emphasis here is not to generate natural-language-like review texts. Since either the original or the created anonymous review will be represented as a vector of independent words in the BOW model, the grammatical requirement is not as strict as that in human languages.

<sup>2</sup>In our experiments, we extract the antonym dictionary from the WordNet lexicon <http://wordnet.princeton.edu/>.

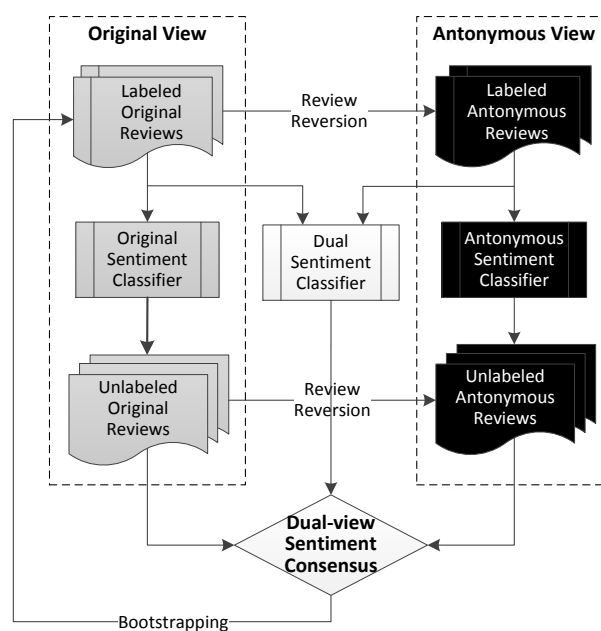


Figure 2: The process of dual-view co-training. Again, the white font color and black background are used to denote the anonymous view.

It is important to notice that the anonymous view removes all negations and incorporates anonymous features. On this basis, we design a dual-view co-training approach. We will introduce our approach in detail in Section 3.2, and analyze its potential advantages in Section 3.3.

### 3.2 The Dual-view Co-training Approach

Since the original and anonymous views form two different views of one review text, it is natural to employ the co-training algorithm, which requires two views for semi-supervised classification.

Co-training is a typical bootstrapping algorithm that first learns a separate classifier for each view using the labeled data. The most confident predictions of each classifier on the unlabeled data are then used to construct additional labeled training data iteratively. Co-training has been extensively used in NLP, including statistical parsing (Sarkar, 2001), reference resolution (Ng and Cardie, 2003), part-of-speech tagging (Clark et al., 2003), word sense disambiguation (Mihalcea, 2004), and sentiment classification (Wan, 2009; Li et al., 2010a).

But it should be noted that the dual views in our approach are different from traditional views. One important property of our approach is that two views are opposite and therefore associated with opposite class labels. Figure 2 illustrates the process of dual-view co-training.

### (1) Dual-view training

For each instance in the initial labeled set, we construct the dual-view representations. Let  $x_o^l$  and  $x_a^l$  denote the bags of words in the original view and the antonymous view, respectively. Note that the class labels in two views are kept opposite:  $y_a^l = 1 - y_o^l$  ( $y \in \{0, 1\}$ ). That is, we reverse the class label in the original view (i.e., positive to negative, or vice versa), as the class label of the created antonymous view.

Suppose  $\mathcal{L}$  is the labeled set, with  $\mathcal{L}_o$  and  $\mathcal{L}_a$  denoting the original-view and antonymous-view labeled sets, respectively. We train two distinct classifiers: the original-view classifier  $h_o$  and the antonymous-view classifier  $h_a$ , based on  $\mathcal{L}_o$  and  $\mathcal{L}_a$ , respectively. We further train a joint classifier by using  $\mathcal{L}_o$  and  $\mathcal{L}_a$  together as the training data, and refer to it as  $h_d$ .

### (2) Dual-view bootstrapping

In standard co-training, we allow each classifier to examine the unlabeled set  $\mathcal{U}$  and select the most confidently predicted examples in each category. The selected examples are then added into  $\mathcal{L}$ , along with the predicted class labels.

In this work, we design a dual-view co-training algorithm to bootstrap the class labels by a joint observation of two sides of one review. Specifically, we propose a new bootstrapping mechanism, based on a principle called **dual-view sentiment consensus**. Given an unlabeled instance  $\{x_o^u, x_a^u\}$ , dual view sentiment consensus requires that, the original prediction  $y_o^u$  and the antonymous prediction should be opposite:  $y_a^u = 1 - y_o^u$ . In other words, we only select the instances of which the original prediction is positive/negative, and the same time the antonymous prediction is negative/positive. To increase the degree of sentiment consensus, we further require that the prediction  $y_d^u$  of  $h_d$  should be the same as  $y_o^u$ .

We sort all unlabeled instances according to the dual-view predictions in each class, filter the list according to the dual-view sentiment consensus principle, and add the top-ranked  $s$  instances in each class to the labeled set. For each selected unlabeled instance, its original view  $x_o^u$  is added into  $\mathcal{L}_o$  with class label  $y_o^u$ ; and the antonymous view  $x_a^u$  is added into  $\mathcal{L}_a$ , with an opposite class label  $y_a^u = 1 - y_o^u$ . When  $\mathcal{L}_o$  and  $\mathcal{L}_a$  receive the supplemental labeled instances, we update  $h_o$  and  $h_a$ .

Our bootstrapping mechanism differs from the traditional methods in two major aspects: First, in traditional co-training, given the same instance, the class labels in two views are the same. But in our approach, the class labels in two views need to be opposite. Second, in traditional co-training, the most confidently predicted examples in each view are selected to extend the amount of labeled data. It is dangerous to believe the confident but incorrect predictions. While in our approach, the candidates are further filtered by the principle of dual-view sentiment consensus. In this way, the labeling accuracy and learning efficiency can be improved.

### (3) Dual-view testing

Finally, in the testing stage, standard co-training uses a joint set of features in two views to train the classifier. In dual-view testing, we use  $h_o$  and  $h_a$  to predict the test example in two views, and make the final prediction by considering both sizes of the review.

Given a test example  $x^{te}$  with its original view denoted by  $x_o^{te}$  and antonymous view denoted by  $x_a^{te}$ , let  $p_o(\cdot|x_o^{te})$  be the posterior probability predicted by the original-view classifier  $h_o$ , and  $p_a(\cdot|x_a^{te})$  be the posterior probability predicted by  $h_a$ . The dual-view testing process can be formulated as follows:

$$p(+|x^{te}) = p(+|x_o^{te}, x_a^{te}) = \frac{p_o(+|x_o^{te}) + p_a(-|x_a^{te})}{2};$$
$$p(-|x^{te}) = p(-|x_o^{te}, x_a^{te}) = \frac{p_o(-|x_o^{te}) + p_a(+|x_a^{te})}{2}.$$

That is, the final positive score is assigned by measuring not only how positive the original review is, but also how negative the antonymous one is; the negative score is assigned by measuring not only how positive the original review is, but also how negative the antonymous one is.

## 3.3 Advantages of Dual-view Co-training

Our proposed dual-view co-training approach has the following three advantages.

### (1) Effectively address the negation issue

We use the antonymous review as a view to effectively address the negation issue. Let us revisit the example in Section 3.1 and assume that the original review (i.e., “The app doesn’t work well on my phone. Disappointing. Do not recommend it.”) is an unlabeled sample. Because the traditional

BOW model cannot well represent negative structures, the review is likely to be incorrectly labeled as positive and then added into the labeled set.

In our proposed approach, the anonymous review (i.e., “The app works well on my phone. Satisfactory. Recommend it.”) removed all the negative structures, and is thus more suited for the BOW representation. In this example, the anonymous review is also likely to be marked as positive. Hence, in this case, both the original review and its anonymous review will be labeled as positive, which violates the principle of dual-view sentiment consensus as mentioned in Section 3.2. As a result, the unlabeled instance will not be added into the labeled set.

Therefore, our approach can overcome the limitations of the conventional methods in addressing the negation issue and reduce the labeling error rate (caused by the negative structures) during the bootstrapping process.

## (2) Automatically learn the associations among antonyms

In semi-supervised sentiment classification, only limited association information between the words and categories can be obtained from a small number of initial labeled data.

For instance, in the above example “disappointing” and “satisfactory” are a pair of antonyms. From the initial labeled data, we may only learn that “disappointing” is derogatory, but we cannot infer that “satisfactory” is commendatory.

During the bootstrapping process in our approach, when constructing the dual view representation, the original view and its anonymous view are required to have opposite class labels. Hence we can automatically infer the relationship between “satisfactory” and “disappointing” (e.g., one is positive and one is negative), thereby improving the learning efficiency of the system.

## (3) Better meet two co-training requirements

Compared with traditional methods, our dual-view co-training can better meet the two co-training requirements: 1) sufficient condition (i.e., each view is sufficient for classification); 2) complementary condition (i.e., the two views are conditionally independent).

First, for the sufficient condition, we use a different view construction method. Most traditional methods construct the two views by feature partitioning (i.e., dividing the original feature set into

two subsets), while we use data expansion by generating anonymous reviews. We will demonstrate in the experimental section (Section 4.6), that our data expansion method can construct better views than the feature partition method in terms of predicting the class labels from individual views.

Second, as we know, every coin has two sides and the two sides are often complementary. In our proposed approach, the original review and its anonymous review (i.e., two sides of one review) are used as two views for co-training and they can better meet the complementary condition. We will illustrate this point in Section 4.6 by calculating the KL divergence between the two views.

## 4 Experimental Study

### 4.1 Datasets and Experimental Settings

We conduct the experiments on the multi-domain sentiment datasets, which were introduced in (Blitzer et al., 2007) and have been widely used in sentiment classification. It consists of four domains (Book, DVD, Electronics, and Kitchen) of reviews extracted from Amazon.com. Each of the four datasets contains 1,000 positive and 1,000 negative reviews. Following the experimental settings used in (Li et al., 2010a), we randomly separate all the reviews in each class into a labeled data set, a unlabeled data set, and a test set, with a proportion of 10%, 70% and 20%, respectively. We report the averaged results of 10-fold cross-validation in terms of classification accuracy.

Note that our approach is a general framework that allows different classification algorithms. Due to the space limitation, we only report the results by using logistic regression<sup>3</sup>. Note the similar conclusions can be obtained by using the other algorithms such as SVMs and naïve Bayes. The LibLinear toolkit<sup>4</sup> is utilized, with a dual L2-regularized factor, and a default tradeoff parameter  $c$ . Similar to (Wan, 2009; Li et al., 2010a), we carry out the experiments with the unigram features without feature selection. Presence is used as the term weighting scheme as it was reported in (Pang et al., 2002) that it performed better than TF and TF-IDF. Finally, the paired  $t$ -test (Yang and Liu, 1999) is performed to test the significance of the difference be-

<sup>3</sup>Logistic regression is quite similar to Maximum Entropy, and has been proved to be more efficient in sentiment classification than some other classification algorithms including naïve Bayes and SVMs (Pang et al., 2002).

<sup>4</sup><http://www.csie.ntu.edu.tw/~cjlin/liblinear/>

	BOOK	DVD	ELEC	KITC	Avg.
Baseline	0.680	0.691	0.726	0.740	0.709
LP	0.681	0.676	0.697	0.722	0.694
T-SVM	0.671	0.677	0.716	0.729	0.698
EM	0.702	0.706	0.758	0.744	0.728
Self-Training	0.689	0.705	0.736	0.751	0.720
Self-Reserved	0.690	0.708	0.735	0.754	0.722
Co-Static	0.696	0.714	0.745	0.762	0.729
Co-Dynamic	0.701	0.725	0.756	0.767	0.737
Co-PI	0.702	0.716	0.746	0.769	0.733
Our approach	<b>0.721</b>	<b>0.738</b>	<b>0.769</b>	<b>0.780</b>	<b>0.752</b>

Table 1: The semi-supervised classification accuracy of ten systems.

tween two systems, with a default significant level of 0.05.

## 4.2 Compared Systems

We implement the following nine systems and compare them with our approach:

- **Baseline**, the supervised baseline trained with the initial labeled data only;
- **Expectation Maximization (EM)**, with the naïve Bayes model proposed by Nigam et al. (2000);
- **Label Propagation (LP)**, a graph-based semi-supervised learning method proposed by Zhu and Ghahramani (2002);
- **Transductive SVM (T-SVM)**, an extension of SVM so that it can exploit unlabeled data in semi-supervised learning (Joachims, 1999);
- **Self-Training**, a bootstrapping model that first trains a classifier, uses it to classify the unlabeled data, and adds the most confident data to the labeled set;
- **Self-Reserved**, a variation of self-training proposed in (Liu et al., 2013), with a reserved procedure to incorporate some less confident examples;
- **Co-Static**, the co-training algorithm by using two static partitions of feature set as two views (Blum and Mitchell, 1998);
- **Co-Dynamic**, a variation of co-training that uses dynamic feature space in each loop. It was reported in (Li et al., 2011) that the Co-Dynamic significantly outperforms Co-Static significantly;
- **Co-PI**, another variation of co-training proposed by (Li et al., 2010a), by using personal

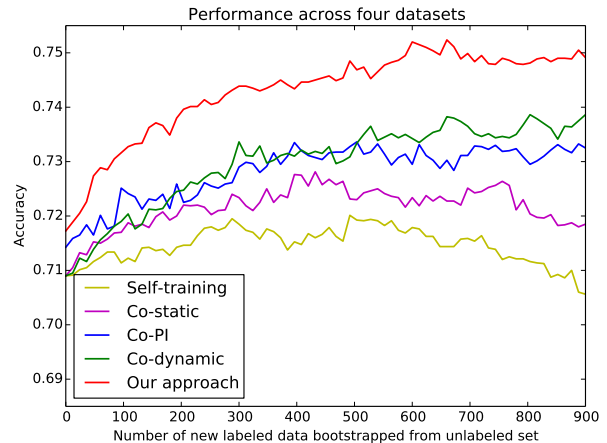


Figure 3: Comparison of different bootstrapping methods.

and impersonal views for co-training.

## 4.3 Performance Comparison

In table 1, we report the semi-supervised classification accuracy of ten evaluated systems. We report the results with 200 labeled, 1400 unlabeled and 400 test reviews. Note that the similar conclusions can be obtained when the size of the initial labeled data changes. We will discuss its influence later.

As can be seen, trained with only 200 labeled data, the supervised baseline yields an average accuracy of 0.709. Self-training gains an improvement of 1.1%. Self-reserved does not show significant priority against Self-training. Three co-training systems (Co-static, Co-dynamic and Co-PI) get significant improvements. They increase the supervised baseline by 2.0%, 2.8% and 2.4%, respectively.

It is somehow surprising that T-SVM and LP do not outperform the supervised baseline, probably because the supervised baseline is obtained by logistic regression, which was reported to be more effective than SVMs in sentiment classification (the supervised result of SVMs is 0.695).

Our proposed approach significantly outperforms all the other methods. It gains the improvement over the supervised baseline, Self-training, Co-static, Co-dynamic and Co-PI by 4.3%, 3.2%, 2.3%, 1.5% and 1.9%, respectively. All of the improvements are significant according to the paired *t*-test.

## 4.4 Comparison of Bootstrapping Methods

In Figure 3, we further compare five bootstrapping methods by drawing the accuracy curve dur-

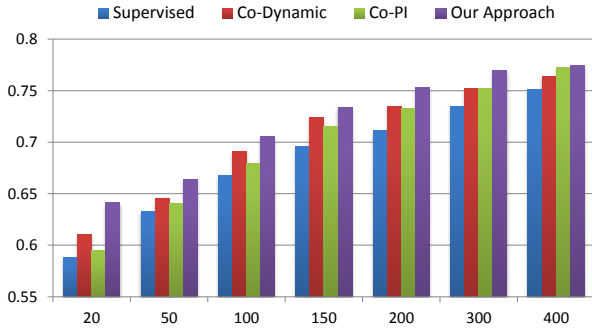


Figure 4: Influence of the size of initial labeled data.

ing the bootstrapping process. The x-axis denotes the number of new labeled data bootstrapped from the unlabeled data.

We can roughly rank five bootstrapping methods as follows: Our approach  $\gg$  Co-dynamic  $>$  Co-PI  $>$  Co-static  $\gg$  Self-training. Self-training gives the worst performance. Co-static works better but the effect is limited. Co-PI and Co-dynamic are significantly better. Our proposed approach outperforms the other systems robustly, along with the increased number of the new labeled data. It suggests that our approach is very efficient in bootstrapping the class labels from the unlabeled data.

#### 4.5 Influence of the Size of the Initial Labeled Set

The above results are obtained with 200 labeled, 1400 unlabeled and 400 test reviews. We now tune the size of the initial labeled set (from 20 to 400), and report its influence in Figure 4. For all the settings, we fix the size of test set as 400. The x-axis denotes the number of initial labeled set. For example, “20” denotes the setting of 20 labeled and 1580 unlabeled data.

We can observe that our all methods improve as the initial size increases. But the improvements become limited when the size becomes larger. When the initial size is 400, the semi-supervised performance is close to the golden result obtained by the supervised classifier trained with all 1600 labeled data.

Our approach performs consistently the best across different sizes of the initial sizes. The smaller the initial size is, the more improvements our approach can gain, in comparison with the other methods. This confirms our analysis in Section 3.3 that the technique of dual-view construction is very effective to boost the semi-supervised

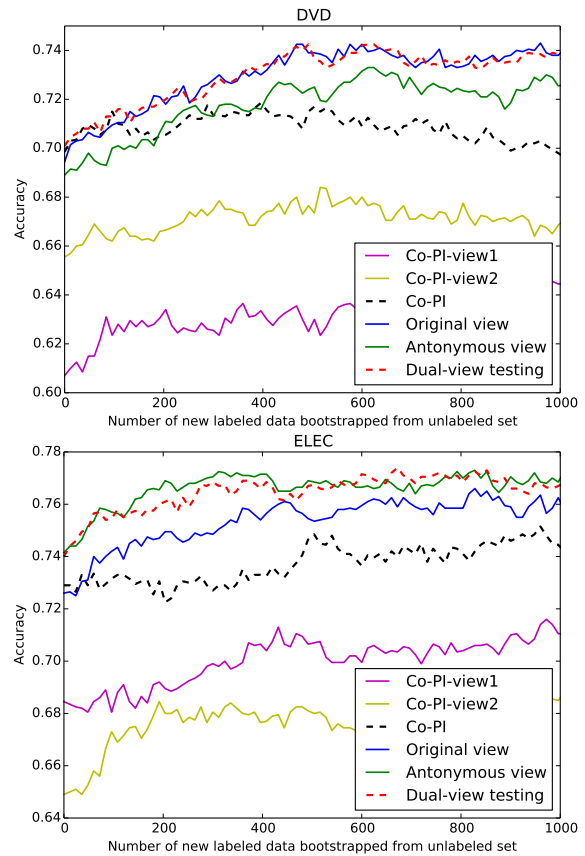


Figure 5: Comparison of different views on the DVD and Electronics datasets.

classification performance, especially when the size of the initial labeled set is small.

#### 4.6 Discussion on the Two Co-training Requirements

Ideally, co-training requires that each view is sufficient for classification (sufficient condition) and two views provide complementary information of the instance, (complementary condition). In this section, we answer the following question empirically: whether our approach could meet the two requirements?

##### (1) Sufficient condition

In Figure 5, we report the classification performance obtained by the classifiers trained with distinct views and compared them with the two views in Co-PI, on the DVD and Electronics datasets. The observation in Book is similar to that in Electronics; the observation in DVD is similar to that in Kitchen.

Seen from Figure 5, the classification performance of both the original-view and antonymous-view classifiers are satisfactory. It shows that in

our approach, each individual view is sufficient to predict the sentiment. In comparison with the two views in Co-PI (i.e., the personal and impersonal views), two views in our approach perform significantly better.

As has been mentioned in Section 3.3, in traditional methods, such as Co-PI and Co-dynamic, two views are created by data partition (or feature partition). In comparison, the two views in our approach are constructed in a manner of data expansion. By creating a new antonymous view, our approach can provide more sufficient information of the reviews than traditional methods.

## (2) Complementary condition

Since we have not found a direct measure of the complementarity of two views, we instead calculate the Kullback-Leibler (KL) divergence between them, based on an assumption that two views with higher KL divergence can provide more complementary information of the instance.

KL divergence is a widely used metric of statistical distance. We assume that distribution of the review text is multinomial, and calculate the K-L divergence between two views as follows:

$$D_{KL}(p||q) = \sum_{i=1}^V p_i \log \left( \frac{p_i}{q_i} \right)$$

where  $p_i$  and  $q_i$  are the probabilities of word appearing in two views, respectively. In our experiments, we use information gain (IG) to select a set of discriminative words with the dimension  $V = 2000$ .

In Table 2, we report the results of three different methods: 1) dataset random partition; 2) personal and impersonal views in Co-PI; 3) original and antonymous views in our approach. We can observe from Table 2 that, random partition has the lowest KL divergence. It shows that the distributional distance between two randomly partitioned views is very small. Co-PI is a higher value, but it still does not have significant difference in two views. By contrast, the KL divergence between the original view and the antonymous view is much higher than both random partition and Co-PI. It demonstrates that the distributions of two views in our approach are significantly different. We thereby infer that the two views constructed in our approach can provide more complementary information than traditional methods. It is reasonable since the antonymous view incorporates the

	KL divergence
Random Partition	2.43
Co-PI	4.59
Our approach	12.33

Table 2: The average KL divergence between two views across four datasets.

antonyms that might have not appeared in the original view (e.g., “satisfactory” in the example in Section 3.2). These features might provide new information about the instance.

## 4.7 The Effect of Dual-view Testing

In Figure 5, we can further observe the effect of dual-view testing. On the Electronics dataset, the antonymous view performs better than the original view. This suggests the advantage of the antonymous view, as it removes the negations and thus is more suitable for the BOW representation. On the DVD dataset, the original view is slightly better. This is also reasonable, because the antonymous review is automatically created and its quality might be limited in some cases. By taking two opposite views into a joint consideration, our dual-view testing technique guarantees a satisfactory classification performance across different datasets.

Note that in the current version, the original-view and antonymous-view classifiers have the same predicting weight. We believe that by learning the tradeoff between two views in different settings may further improve our approach’s performance. For example, if the original view on the Electronics dataset gets a relatively larger weight, dual-view testing might gain more improvements.

## 5 Conclusions

In this work, a review text is represented by a pair of bags-of-words with opposite views (i.e., the original and antonymous views). By making use of two views in pairs, a dual-view co-training algorithm is proposed for semi-supervised sentiment classification. The dual-view representation is in a good accordance with the two co-training requirements (i.e., sufficient condition and complementary condition). The experimental results demonstrate the effect of our approach, in addressing the negation problem and enhancing the bootstrapping efficiency for semi-supervised sentiment classification.



## Acknowledgement

The work is supported by the Natural Science Foundation of China (61305090), and the Jiangsu Provincial Natural Science Foundation of China (BK2012396).

## References

- A. Aue and M. Gamon. 2005. Customizing Sentiment Classifiers to New Domains: A Case Study. In *Proceedings of Recent Advances in Natural Language Processing*.
- J. Blitzer, M. Dredze, and F. Pereira. 2007. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. In *Proceedings of ACL*.
- A. Blum and T. Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of COLT*.
- S. Clark, J. R. Curran, and M. Osborne. 2003. Bootstrapping POS taggers using unlabelled data. In *Proceedings of CoNLL*.
- S. Dasgupta and V. Ng. 2009. Mine the Easy and Classify the Hard: Experiments with Automatic Sentiment Classification. In *Proceedings of ACL-IJCNLP*.
- A. Goldberg and X. Zhu. 2006. Seeing stars when there aren't many stars: graph-based semi-supervised learning for sentiment categorization. In *Proceedings of the Workshop on TextGraphs at HLT-NAACL*.
- M. Hu and B. Liu. 2004. Mining opinion features in customer reviews. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*.
- D. Ikeda, H. Takamura, L. Ratinov, and M. Okumura. 2008. Learning to Shift the Polarity of Words for Sentiment Classification. In *Proceedings of IJCNLP*.
- T. Joachims. 1999. Transductive Inference for Text Classification using Support Vector Machines. In *Proceedings of ICML*.
- A. Kennedy and D. Inkpen. 2006. Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*, 22:110–125.
- LaTeX Error: File 'url.sty' not found. V. Ng and C. Cardie. 2003. Weakly supervised natural language learning without redundant views. In *Proceedings of HLT-NAACL*.
- K. Nigam, A. McCallum, S. Thrun, and T. Mitchell. 2000. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2/3): 103–134.
- S. Li, C. Huang, G. Zhou, and S. Y. M. Lee. 2010a. Employing personal/impersonal views in supervised and semi-supervised sentiment classification. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.
- S. Li, S. Lee, Y. Chen, C. Huang, and G. Zhou. 2010b. Sentiment Classification and Polarity Shifting. In *Proceeding of the International Conference on Computational Linguistics (COLING)*.
- S. Li, Z. Wang, G. Zhou, and S. Lee. 2011. Semi-Supervised Learning for Imbalanced Sentiment Classification. In *Proceedings of IJCAI*.
- S. Li. 2013. Sentiment classification using subjective and objective views. *International Journal of Computer Applications*, 80(7): 30–34.
- Z. Liu, X. Dong, Y. Guan, and J. Yang. 2013. Reserved Self-training: A Semi-supervised Sentiment Classification Method for Chinese Microblogs. In *Proceedings of IJCNLP*.
- R. Mihalcea. 2004. Co-training and self-training for word sense disambiguation. In *Proceedings of CoNLL*.
- J. Na, H. Sui, C. Khoo, S. Chan, and Y. Zhou. 2004. Effectiveness of simple linguistic processing in automatic sentiment classification of product reviews. In *Proceeding of the Conference of the International Society for Knowledge Organization*.
- S. Orimaye, S. Alhashmi, and E. Siew. 2012. Buy it - don't buy it: sentiment classification on Amazon reviews using sentence polarity shift. In *Proceedings of the Pacific Rim International Conferences on Artificial Intelligence (PRICAI)*.
- B. Pang, L. Lee, and S. Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of EMNLP*.
- L. Polanyi and A. Zaenen. 2004. Contextual lexical valence shifters. In *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text*.
- Y. Ren, N. Kaji, N. Yoshinaga, M. Toyoda, and M. Kitsuregawa. 2011. Sentiment Classification in Resource-Scarce Languages by using Label Propagation. In *Proceedings of the Pacific Asia Conference on Language, Information and Computation (PACLIC)*.
- A. Sarkar. 2001. Applying cotraining methods to statistical parsing. In *Proceedings of NAACL*.
- R. Socher, J. Pennington, E. H. Huang, and A. Y. 2012. Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions. In *Proceedings of EMNLP*.
- Y. Su, S. Li, S. Ju, G. Zhou and J. Li. 2012. Multi-view Learning for Semi-supervised Sentiment Classification. In *Proceedings of the International Conference on Asian Language Processing*.

- X. Wan. 2009. Co-Training for Cross-Lingual Sentiment Classification. In *Proceedings of ACL-IJCNLP*.
- R. Xia, T. Wang, X. Hu, S. Li, and C. Zong. 2013. Dual Training and Dual Prediction for Polarity Classification. In *Proceedings of ACL*.
- Y. Yang and X. Liu. 1999. A re-examination of text categorization methods. In *Proceedings SIGIR*.
- S. Zhou, Q. Chen, and X. Wang. 2010. Active Deep Networks for Semi-Supervised Sentiment Classification. In *Proceedings of COLING*.
- X. Zhu and Z. Ghahramani. 2002. Learning from labeled and unlabeled data with label propagation. *Technical Report CMU-CALD-02-107*, Carnegie Mellon University.
- X. Zhu, Z. Ghahramani, and J. Lafferty. 2003. Semi-supervised learning using Gaussian fields and harmonic functions. In *Proceedings of the International Conference on Machine Learning (ICML)*.