SELC: A Self-Supervised Model for Sentiment Classification

Likun Qiu[1,2], Weishi Zhang[2,3], Changjian Hu[2], Kai Zhao[2]
[1]Department of Chinese Language and Literature, Peking University, China
[2]NEC Laboratories, China
[3]School of Software, Tsinghua University, China
{qiulikun, zhangweishi, huchangjian, zhaokai}@research.nec.com.cn

ABSTRACT

This paper presents the SELC Model (**SE**lf-Supervised, **L**exicon-based and **C**orpus-based Model) for sentiment classification. The SELC Model includes two phases. The first phase is a lexicon-based iterative process. In this phase, some reviews are initially classified based on a sentiment dictionary. Then more reviews are classified through an iterative process with a negative/positive ratio control. In the second phase, a supervised classifier is learned by taking some reviews classified in the first phase as training data. Then the supervised classifier applies on other reviews to revise the results produced in the first phase. Experiments show the effectiveness of the proposed model. SELC totally achieves 6.63% F_1-score improvement over the best result in previous studies on the same data (from 82.72% to 89.35%). The first phase of the SELC Model independently achieves 5.90% improvement (from 82.72% to 88.62%). Moreover, the standard deviation of F_1-scores is reduced, which shows that the SELC Model could be more suitable for domain-independent sentiment classification.

Categories and Subject Descriptors

H.3.1 [**Information Systems**]: INFORMATION STORAGE AND RETRIEVAL –*Content Analysis and Indexing.*

General Terms

Algorithms, Performance, Experimentation.

Keywords

Information retrieval, opinion mining, sentiment classification.

1. INTRODUCTION

There are a lot of product reviews on the Web. In those reviews, people evaluate the products they used before and express their feelings about the products. The analysis of reviews is helpful for both consumers and the producers. In particular, assigning the positive and negative sentiment values to product reviews is referred to as sentiment classification. Here, the polarities include positive and negative polarities.

Generally, there are two types of approaches tackling the sentiment classification task according to the knowledge they used. One is corpus-based [2, 3, 6, 9, 10] and the other is lexicon-based [4, 13, 15, 16]. The corpus-based approaches are usually supervised, i.e., requiring training sets, and performing well when the training set is large enough and correctly labeled. On the contrary, the lexicon-based approaches are mostly unsupervised, requiring only a few seed words such as "excellent" or "poor". Compared with the corpus-based approaches, lexicon-based ones have the advantage of domain-independence because it is much easier to acquire a few seed words than correctly label a large training set, and some researchers even showed that the seed words can be automatically generated [16].

Further study [1] showed that corpus-based and lexicon-based approaches could be complementary to each other. In particular, corpus-based ones usually achieve higher precision on positive reviews while lexicon-based ones usually achieve higher precision on negative reviews. In other words, corpus-based approaches tend to classify a review as negative while lexicon-based approaches tend to positive. The two tendencies are referred to as negative classification bias and positive classification bias respectively. The integration of the two types of approaches seems very promising. However, the direct integration like [1] is a kind of supervised approach in essence because it still needs a small-scale manually annotated corpus.

We attempt to build a model to have the following features:

- be domain-independent, which means nothing of the model needs to be changed when the domain changes;
- exploit the complementarities between lexicon-based and corpus-based approaches to improve the whole performance;
- do not need to manually annotate corpus in the integration process.

The SELC Model is proposed to meet the above requirements. It is built based on a lexicon-based approach. Therefore it is domain independent. Several innovations, including positive/negative ratio control, the use of a general sentiment dictionary and the enlargement of negation word list, are adopted to overcome the positive classification bias of lexicon-based approaches. Then the SELC Model introduces a corpus-based approach to revise the

CIKM'09, November 2–6, 2009, Hong Kong, China.

result of the lexicon-based one. In the revising process, the corpus-based approach takes some reviews classified by the lexicon-based one as training data. Although the training data is machine-generated, most reviews of the training data are correctly labeled (above 93% precision in the experiments), because the lexicon-based approach is well designed. As such, the performance of the corpus-based approach is still reliable, and the whole performance is improved.

The SELC Model achieves an overall F_1-score of 89.35% on data sets of ten domains, with an improvement of 6.63% over the best result (82.72%) in previous studies on the same data [16].

The rest of this paper is organized as follows. Section 2 surveys related work. The overview of our approach is presented in Section 3. Then Section 4 and 5 describe the details of the SELC Model. Experiments are described in Section 6. Section 7 discusses the experiment results and gives an error analysis. The final section gives conclusions and proposes future work.

2. RELATED WORK

Document sentiment classification methods might be classified into corpus-based methods and lexicon-based methods according to the knowledge they used. The integration of these two methods is a new direction.

2.1 Corpus-based Methods

Most corpus-based sentiment classifiers use standard machine learning techniques such as SVM and NBm [5]. Different factors affecting the machine learning process were investigated. For instance, linguistic, statistical and n-gram features are used in [10]. Semantically oriented words are used to identify the polarity at the sentence level in [6]. A graph-based technique is used in [3] to identify and analyze only subjective parts of texts. Selected words and negation phrases were investigated in [8]. Such approaches work well in situations where large labeled corpora are available for training.

But the performance of corpus-based methods generally decreases when training data is insufficient or acquired from a different domain [2] [9], topic [9] or time period [9].

To solve that problem, unsupervised and weakly supervised methods can be used to take advantage of a small number of annotated in-domain examples and/or unlabelled in-domain data. For instance, the method of [2] trains a model on a small number of labeled examples and large quantities of unlabelled in-domain data. In [11], structural correspondence learning is applied to the task of domain adaptation for sentiment classification of product reviews. Similarly, the authors of [14] suggested combining out-of-domain labeled examples with unlabelled ones from the target domain in order to solve the domain-transfer problem. So far, the performance of such methods is inferior to the supervised approaches with in-domain training.

2.2 Lexicon-based Methods

Lexicon-based methods are usually unsupervised. Some of them use general sentiment word lists acquired from the Internet or dictionaries. It is showed in [4] that lexicon-based methods perform worse than statistical models built on sufficiently large training sets in the movie review domain. [12] shows that the performance of systems using general word lists is comparable to that of supervised machine learning approaches on some domains such as product reviews.

Other methods use seed words to replace the word list, and then enrich the seed words by more sentiment words and phrases. For instance, the method in [13] uses two human-selected seed words (*poor* and *excellent*) in conjunction with a very large text corpus. The semantic orientation of phrases is computed as their association with the seed words (measured by point-wise mutual information). The sentiment of a document is calculated as the average semantic orientation of all such phrases. In [16], seed words are automatically generated based on a linguistic pattern, which is called *negated adverbial construction*. Experimental results show that this method achieves similar performance to supervised methods.

2.3 Integration of Corpus-based Methods and Lexicon-based Methods

In [1], a corpus-based classifier trained on a small set of annotated in-domain data is integrated with a lexicon-based system trained on *WordNet*. The experiments show that the hybrid method brings significant gains in accuracy and recall over both the individual corpus-based and lexicon-based method. Although the approach is very promising, it still requires a certain amount of annotated corpus and therefore a supervised method in essence. Moreover, their experiments are conducted on a corpus with equal number of positive and negative examples. The effectiveness of their approach on unbalanced dataset still needs to be verified.

3. OVERVIEW OF OUR APPROACH

3.1 Corpus-based methods VS Lexicon-based methods

The complementation property of corpus-based and lexicon-based methods, i.e., one classifier makes an error, while the other one gives the correct answer, was initially exploited in [1]. However, the reason behind this phenomenon was not revealed. Here we attempt to give an explanation.

In many cases, people are accustomed to directly use positive or negative words to express their positive or negative sentiment. This is referred to as *direct sentiment expression*. However, in many other cases, people convey positive feeling with negative words and convey negative feeling with positive words, with the help of negative constructions. This is referred to as *indirect sentiment expression*. Between the two kinds of indirect sentiment expression, conveying negative feeling with positive words in negative constructions is more popular. This is referred to as *indirect expression of negative sentiment* (IENS). For instance, in most cases, people say 不好 bu-hao '*not good'* to express unsatisfactory feeling and say 不太高 bu-tai-gao '*not very tall'* to convey similar meaning with 矮 *ai* 'short'. As shown in Table 1, the frequency of IENS (3,554), is much higher than that of the indirect expression of positive sentiment in Chinese (616) and is very close to that of the direct expression of negative sentiment.

Table 1 Distribution of Sentiment Words in Chinese[1]

Documents	Words	Separate	With Negation
All	Positive	19914	**3554**
	Negative	4642	616
Positive	Positive	15958	**1040**
	Negative	1250	320
Negative	Positive	3956	**2514**
	Negative	3392	296

The popularity of IENS is just the reason of the former difference between corpus-based methods and lexicon-based methods. In lexicon-based methods, the polarities of words are assigned in a dictionary in advance. IENS means that positive words are used frequently even in negative documents but negative words are scarcely used in positive documents. Therefore, positive words sometimes predominate even in negative documents, and therefore lexicon-based methods are apt to classify a document as positive. To improve the result of lexicon-based methods, more constraints for classifying a document as positive should be added. For instance, the negation constructions used in [16] is such kind of constraints.

In corpus-based methods, the polarities of words are learned automatically by machine learning methods. Since both negative words and positive words might be frequently used to convey negative sentiment, yet only positive words are frequently used to convey positive sentiment, it is easy for a corpus-based method to learn negative expressions. Therefore, corpus-based methods are apt to classify a document as negative.

3.2 Overview of Our Approach

The SELC Model is proposed to exploit the complementarities between lexicon-based and corpus-based methods to improve the whole performance. This model consists of two phases. In Figure 1, Phase 1 blocks are grouped in the solid-line frame and Phase 2 blocks in the dash-line frame.

Phase 1 is a lexicon-based iterative process. In this phase, a sentiment vocabulary is initialized by a general sentiment dictionary. The vocabulary is used to classify reviews. Then more sentiment words are found from the classified reviews and update the vocabulary. The new vocabulary then helps classify more reviews. By this iterative process, the vocabulary and classified reviews are updated (and generally enlarged) step by step. In the iterative process, a new technology, i.e., positive/negative ratio control, is introduced. That control ranks the reviews and keeps the same number of top-ranked positive and negative reviews in each iteration. Because of the ratio control, the iterative process completes when the set of classified positive or negative reviews does not enlarge any more. All the reviews having been classified at this point form the *Classified Set* (part A in Figure 1). There are still some reviews left unclassified, which compose the *Uncertain Set*. Then current vocabulary is used to classify the reviews of *Uncertain Set* without ratio control. No iteration either. The result is marked as part B in Figure 1.

Phase 2 is used to integrate the results of a corpus-based supervised model and the results of Phase 1. In Phase 2, the supervised model (SVM in Figure 1) takes the *Classified Set* as training data. The achieved model is then used to classify the reviews of *Uncertain Set*. The supervised model only applies on the *Uncertain Set*, but not the whole set of reviews, because the *Uncertain Set* is not classified as well as those reviews in the *Classified Set*, where ratio control makes a strict restriction and therefore keeps a high precision. That is also the reason of the feasibility to take *Classified Set* as training data for a supervised model, even though the data is machine-classified, but not human-labeled. Finally, the results of the corpus-based model are integrated with the results of Phase 1 (part C of Figure 1).

Phase 1 can be used independently. It is referred to as the Basic SELC Model. The result of the Basic SELC Model is A and B part in Figure 1. Accordingly, the complete model is referred to as the Standard SELC Model, which is abbreviated as the SELC Model. The result of the SELC Model is A and C part in Figure 1.

The details of the two phases are described in the following two sections.

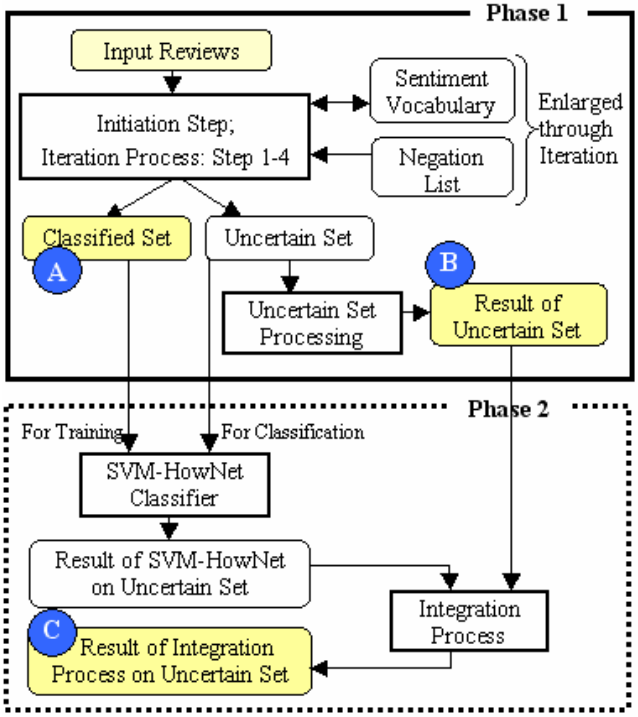

Figure 1 Flow Chart of The SELC Model

Results of The Basic SELC Model = A + B

Results of The SELC Model = A + C

4. PHASE 1

Based on a Chinese sentiment dictionary and a negation list, Phase 1 of the SELC Model uses an iterative process to enlarge the sentiment vocabulary and improve overall accuracy mutually and gradually. Phase 1 includes six steps, which are introduced in the following.

[1] The statistical data are got from 7,792 Chinese product reviews.

4.1 Initiation Step

The sentiment vocabulary, denoted by V_{sen}, includes a list of items, each of which is assigned with a sentiment score. V_{sen} is initialized by a sentiment dictionary, which usually includes a lot of positive and negative words. A positive word is assigned with score +1, while a negative word is assigned with score -1. Monosyllabic words are filtered from V_{sen} because most of them are too ambiguous to provide reliable sentiment.

4.2 Step 1: Computation of Review Sentiment Scores

First, each review is divided into zones by punctuation marks.

Second, for a zone, each item in current V_{sen} is checked occurring in the zone or not. If an item does occur in the zone, it is taken as an effective item of the zone.

Third, each effective item of a zone is scored by Equation (1), where L_d is the length of the item, L_{phrase} is the length of the current zone, S_d is the item's score in current V_{sen}, and N_d is a negation check coefficient with a default value of 1. If the lexical item is preceded by a negation within the current zone, N_d is set to -1.

$$S_i = \frac{L_d^{\,2}}{L_{phrase}} S_d N_d \quad (1)$$

Then the scores of all the effective items of a zone are summed up as ZoneScore (the sentiment score of the zone). If the ZoneScore of a zone is greater than zero, it is classified as positive. If the ZoneScore of a zone is smaller than zero, it is classified as negative. Otherwise, it is left unclassified.

Finally, the scores of all zones in each review are summed up as ReviewScore (the sentiment score of a review).

4.3 Step 2: Review Sentiment Classification with Ratio Control

1 Let C_{min}=MIN($C_{positive}$, $C_{negative}$).
2 Sort all reviews in descending order by their ReviewScores assigned in Step 1.
3 Tagging:
3.1 Tag the former C_{min} reviews as positive.
3.2 Tag the latter C_{min} reviews as negative.
3.3 Others are left unclassified.

Figure 2 Review Sentiment Classification with Ratio Control

Basically, a review is classified as positive if its ReviewScore is greater than zero, or negative if its ReviewScore smaller than zero. This policy looks good but would cause sentiment bias for items.

In each iteration process, since there are generally different numbers of positive and negative reviews, the strength of sentiment polarity of items may be biased. For example, if there are 20 positive reviews and 10 negative reviews, and the word 'I' occurs in all the 30 reviews, it will have a sentiment score of 10, and be judged as a positive item. But in fact, such a word may have no sentiment polarity. To overcome the bias caused by unequal number of positive and negative reviews, a ratio control is introduced. It requires the numbers of positive and negative reviews classified in each iteration to be the same. Denote the number of reviews with a positive ReviewScore as $C_{positive}$ and the number of reviews with a negative ReviewScore as $C_{negative}$. Then, review sentiment classification with ratio control is realized in the following way (see Figure 2).

4.4 Step 3: Iterative Retraining

The sentiment vocabulary V_{sen} is updated (and usually enlarged) in this step. Each lexical item[2] that occurs at least twice in those classified reviews is taken as a candidate item. The difference between each candidate item acting as a positive item and a negative item are measured by Equation (2). F_p and F_n denote the frequencies of the candidate item in positive reviews and negative reviews respectively. If the item is preceded by a negation in current zone and the current review is positive, its corresponding frequency F_p is reduced by one, or vice versa. Therefore, the value of F_p and F_n might be a negative number.

$$difference = \frac{\left|F_p - F_n\right|}{\left(F_p + F_n\right)} \quad (2)$$

Only when the difference is evident enough, the candidate item can be added into V_{sen}. The threshold is set as 1. That is, if its difference score is not less than 1, it can be added into V_{sen}. Notice that those items occurring only in positive (or negative) reviews can be included in V_{sen}, as their difference score is 1.

Then the sentiment score of each item in V_{sen} is recalculated according to Equation (3).

$$F_p - F_n \quad (3)$$

4.5 Step 4: Iteration Control

This step is used to determine when the iterative process completes. If there is no difference in the classification result between two iterations, the iterative process completes. When the iteration process completes, the system goes to the Uncertain Set Processing Step. Otherwise, it goes to Step 1 and a new iteration starts.

4.6 Uncertain Set Processing Step

For an unclassified review R,
1 If $ZC_{positive}$ > $ZC_{negative}$, R is tagged positive.
2 Else if $ZC_{positive}$ < $ZC_{negative}$, R is tagged negative.
3 Else:
3.1 If ReviewScore > 0, R is positive.
3.2 Else if ReviewScore < 0, R is negative.
3.3 Else R is left unclassified.

Figure 3 Uncertain Set Processing Step

The ratio control requires the numbers of negative and positive documents to be equal in the iteration process. Therefore, when the iteration retraining completes, there are still a few reviews left unclassified. Denote the number of positive zones in a document as $ZC_{positive}$ and the number of negative zones in a review as $ZC_{negative}$. Denote the review to be classified as R. Then, the

[2]Let N be the length of a zone, a lexical item is a sequence of Chinese characters excluding punctuation marks, from unigram to N-gram, in an enclosing zone.

reviews of Uncertain Set are classified in the following way (see Figure 3).

5. PHASE 2

5.1 Corpus-based Supervised Method

We choose SVM as the machine-learning method to implement the corpus-based supervised method. The model uses a general sentiment dictionary as the feature set. TFIDF measure (see Equation (4)) is used to compute weights.

$$w_i = tf_i \times \log \frac{N_i}{df_i} \quad (4)$$

5.2 Integration Process

This phase is specially designed to process reviews in the *Uncertain Set*. In this phase, the results of the corpus-based model (CB result) and the results of the first phase (LB result, Lexicon-Based result) are integrated together.

As mentioned in Section 3, the lexicon-based model classifies most reviews and left some reviews as the *Uncertain Set* when the iterative process completes. The corpus-based model takes those classified reviews as training data and those reviews in the Uncertain Set as test data. Because of the positive classification bias of lexicon-based methods and the negative classification bias of corpus-based methods (see Section 1), lexicon-based methods usually can achieve high precision on negative reviews while corpus-based methods usually can achieve high precision on positive reviews. Therefore, the two kinds of results are integrated in the following way (see Figure 4).

Given a review,

1. **If the two results were the same, they would be taken as the final result;**
2. **Else if the LB result were negative, it would be taken as the final result;**
3. **Else the CB result would be taken as the final result.**

Figure 4 Integration Process

6. EXPERIMENTS

6.1 Data and Tools

The experiments in this paper were conducted on the data sets of 7,779 product reviews written in Chinese. All the reviews concern with ten domains (sub-corpora [3]): Monitors, Mobile phones, Digital Cameras, MP3 players, Computers parts, Video cameras and lenses, Networking, Office equipment, Printers, Computer peripherals. In the following, they are indexed as C1 to C10 respectively. Each sub-corpus has equal number of positive and negative reviews.

For all the experiments in this paper, the *HowNet* Sentiment Dictionary[4] is used as the sentiment dictionary. The dictionary contains 4566 positive words and 4370 negative words.

WEKA 3.4.11 (http://www.cs.waikato.ac.nz/~ ml/weka) is used as the implementation of Support Vector Machine (SVM) classifiers.

6.2 Baseline

The result reported in [16] (see left column of Table 2) is taken as the baseline of SELC. The overall F_1-score is 82.72% in [16], and the standard deviation of the F_1-scores on ten sub-corpora is 5.22%.

6.3 Results of the Basic SELC Model

In Step 1 and Step 3 of Phase 1, an enlarged negation word list containing ten negations is used: {不 bu '*not*', 不会 bu-hui '*would not*', 没有 mei-you '*don't have*', 没 mei '*don't have*', 虽然 sui-ran '*although*', 虽 sui '*although*', 尽管 jin-guan '*although*'', 缺 que '*don't have*', 缺乏 que-fa '*don't have*', 无 wu '*don't have*'}.

The results of the Basic SELC Model are shown in the right column of Table 2. It achieves an overall F_1-score of 88.62%, which improves 5.90% over the baseline. The standard deviation of the F_1-scores on ten sub-corpora is only 2.35%, which improves 2.87% over the baseline. The drop on the standard deviation shows that the Basic SELC Method is more domain-independent than the baseline. Moreover, the average iteration number of the Basic SELC Model is 5.6, with a decrease of 5.9 over that of [16] (11.5).

Table 2 Comparison between baseline and Basic SELC Model

	Baseline [16]			**Basic SELC Model**		
	P	**R**	**F_1**	**P**	**R**	**F_1**
C1	85.57	85.07	85.32	90.38	89.46	89.92
C2	92.63	92.19	92.41	90.54	89.68	90.11
C3	84.92	83.58	84.24	87.66	85.87	86.76
C4	88.69	87.55	88.11	91.71	90.89	91.30
C5	77.78	77.27	77.52	86.14	84.74	85.43
C6	83.62	81.99	82.80	89.11	88.37	88.73
C7	72.83	72.00	72.41	84.20	83.71	83.95
C8	82.42	81.34	81.88	89.13	88.54	88.83
C9	81.04	79.61	80.32	90.27	89.63	89.95
C10	82.24	82.06	82.15	91.43	91.03	91.23
AVG	83.17	82.27	**82.72**	89.06	88.19	**88.62**

As mentioned in Section 4.7, there are several novel aspects in the Basic SELC Model, which affect the performance simultaneously. To check their individual effect, three variant models were implemented. They are referred to as V1, V2 and V3 respectively. In V1, the ratio control is removed. In V2, the *HowNet* Sentiment Dictionary is replaced by the seed set automatically generated in [16]. In V3, the ten negations are replaced by the six negations of [16]. Table 3 shows that both the ratio control and the *HowNet* Sentiment Dictionary take great effect on the performance, i.e., improving 13.40% and 6.52% F_1-score respectively. The enlargement of negation word list also improves 1.14% F_1-score.

[3] It is provided by Zagibalov and Carroll (http://www.informatics.sussex.ac.uk/users/tz21/coling08.zip)

[4] http://www.keenage.com/download/sentiment.rar

Table 3 The Results of Three Variants of the Basic SELC Model

	F_1			
	V1	**V2**	**V3**	**Best**
C1	77.16	86.30	90.25	89.92
C2	82.25	86.29	89.33	90.11
C3	75.25	84.20	87.57	86.76
C4	74.91	88.07	89.75	91.30
C5	69.05	73.38	81.01	85.43
C6	75.00	84.22	86.77	88.73
C7	69.74	65.04	84.68	83.95
C8	73.86	85.88	87.03	88.83
C9	75.11	82.38	88.53	89.95
C10	79.91	85.23	89.90	91.23
AVG	**75.22**	**82.10**	**87.48**	**88.62**

6.4 Results of the SELC Model

We implement a classifier based on SVM, taking the words in *HowNet* Sentiment Dictionaries as features. This is referred to as SVM-*HowNet* classifiers. The Basic SELC Model without the Uncertain Set Processing Step is referred to as the Basic SELC* Model ("A" part of Figure 1). The results of the Basic SELC* Model, the Basic SELC Model and SVM-*HowNet* classifier on negative and positive reviews are shown in Table 4. The Basic SELC* Model achieves similar precisions on negative and positive reviews (93.51% and 92.95%). It shows that the novel technologies in this model have effectively overcome the positive classification bias of lexicon-based methods. However, for the Basic SELC Model, the difference of precision is still evident between on negative reviews (93.14%) and positive reviews (85.71%). The bias is mainly caused by the processing of *Uncertain Set*.

The results of the SVM-*HowNet* classifier in 10-fold cross-validation mode are also shown in Table 4. Those results fit for the analysis in Section 1. That is, a corpus-based classifier tends to classify a review as negative (with higher precision on positive reviews (88.96%) and higher recall on negative reviews (89.43%) respectively).

Table 5 shows the results of the Basic SELC Model, the SVM-*HowNet* classifier and their integrated results on Unclassified Set. It shows that the integrated results improve 7.71% and 12.15% over the SVM-classifier and the Basic SELC Model respectively. The improvement is appreciable. However, since the SVM-classifier only applies on the *Uncertain Set*, which only takes a small proportion (about 9%) of the entire set, the overall improvement of the SELC Model is not so prominent.

The SELC Model achieves an overall F_1-score of 89.35% (see Table 6), with an improvement of 6.63% over the baseline (82.72% of [16]) and 0.73% over the Basic SELC Model (88.62% in Table 2).

Table 4 Results of the Basic SELC* Model, the Basic SELC Model and SVM-*HowNet* Classifier on Negative and Positive Reviews

Model	**Document Sentiment**	**P**	**R**	**F_1**
Basic SELC*	Pos	92.87	93.51	93.18
	Neg	93.56	92.95	93.25
	All	93.22	93.23	93.22
Basic SELC	Pos	85.71	92.73	89.07
	Neg	93.14	83.66	88.12
	All	89.06	88.19	88.62
SVM-*HowNet*	Pos	88.96	84.74	86.80
	Neg	85.43	89.43	87.38
	All	87.20	87.09	87.14

Table 5 Results of the Basic SELC Model, SVM-*HowNet* Classifier and Their Integrated Results on Uncertain Set

Model	**P**	**R**	**F_1**
Basic SELC	61.26	41.57	**49.44**
SVM-*HowNet*	53.88	53.88	**53.88**
Integrated	61.59	61.59	**61.59**

Table 6 Results of the SELC Model

	P	**R**	**F_1**
C1	90.78	90.78	90.78
C2	90.20	90.20	90.20
C3	87.21	87.21	87.21
C4	92.17	92.17	92.17
C5	86.69	86.69	86.69
C6	90.03	90.03	90.03
C7	84.00	84.00	84.00
C8	90.18	90.18	90.18
C9	90.51	90.51	90.51
C10	91.68	91.68	91.68
AVG	89.35	89.35	**89.35**

6.5 Results of the Basic SELC Model and the SELC Model on Corpora with Different Ratios between Positive and Negative Reviews

The corpus used in the above experiments consists of half positive documents and half negative documents, i.e., the positive/negative ratio is 1:1. That is the usual way of corpus construction [1, 4, 7 16]. But that is not always the state of real-world data, in which positive reviews might predominate and even take a proportion of more than 80% [16].

Therefore, corpora with positive/negative ratio of 6:4, 7:3 and 8:2 are constructed respectively based on the 1:1 corpus. For instance, in the 6:4 case, one third of the negative reviews are randomly selected and removed. The results (Table 7) show that both the

Basic SELC Model and HUCL Model perform consistently well on the three cases (with an F_1-score between 86.68% and 89.70%).

Table 7 Results of the Basic SELC Model and the SELC Model on Corpora with Different Ratios between Positive and Negative Reviews

Pos/Neg Ratio	Model	P	R	F_1
5:5	Basic SELC	89.06	88.19	88.62
	SELC	89.35	89.35	89.35
6:4	Basic SELC	89.75	88.81	89.28
	SELC	89.33	89.33	89.33
7:3	Basic SELC	90.16	89.24	89.70
	SELC	89.00	89.00	89.00
8:2	Basic SELC	87.50	86.75	87.12
	SELC	86.68	86.68	86.68

7. DISCUSSION AND ERROR ANALYSIS

7.1 Discussion

A set of factors involved in the Basic SELC Model enables substantial performance gains. All those factors contribute to overcome the positive classification bias of lexicon-based methods. First, the ratio control is introduced into the iteration process. The ratio control can help balance the negative/positive items in the iteration process. For instance, if the ratio of positive documents is too large, accordingly, the ratio of positive items will also increase. In such a case, the ratio control can decrease the increasing speed of positive words and therefore help overcome the positive classification bias of lexicon-based methods.

Second, in the Initiation Step, a general sentiment dictionary is used to replace a seed set generated automatically (see [16]). A seed set usually contains a small number of words while a dictionary contains a lot of words. By introducing more sentiment words in the initial step, less error is generated in the initiation step and propagated in the following iterations. In addition, a seed set may not balance between positive and negative words. For example, the seed set generated in [16] only contains positive words but no negative words. Thus, negation words like 不 bu '*not*' are relied on to judge negative reviews. However, negation words are ambiguous sometimes. For example, 不知道是否清楚 bu-zhi-dao-shi-fou-qing-chu '*do not know whether it is clear or not*' is not the same as 不清楚 bu-qing-chu '*not clear*'. Since a general dictionary contains a lot of negative words, the dependence on negation words is much decreased.

Third, among the ten negations, 没有, 缺, 缺乏 and 无 have similar meaning with 没. 虽然, 虽 and 尽管 are conjunctions functioning similarly to negations like 不 in the sense that all of them transform the sentiment of the following words to their opposite. Six negations are used in [16]: {不, 不会, 没有, 摆脱 bai-tuo '*get rid of*', 免去 mian-qu '*excuse*', 避免 bi-mian '*avoid*'}. But only three of them, 不, 不会 and 没有, are negation words. The remaining three words 摆脱, 免去 and 避免 usually directly convey negative sentiment, but not transform the sentiment of their following words. As the negation list is enlarged properly, the chances of classifying a review as negative increases accordingly. Therefore, this change also helps overcome the positive classification bias of lexicon-based methods.

In Phase 2, the utilization of supervised method further improves the overall performance. Although the training data used in Phase 2 is tagged automatically in Phase 1, the supervised method performs effectively because of the high precision of the result of Phase 1.

7.2 Error Analysis

Most of the errors are caused by ambiguous sentiment words such as 多 duo '*many*' and 少 shao '*few*'. The sentiments of those words usually vary within different contexts. For instance, 优点多 you-dian-duo '*many advantages*' conveys positive sentiment but 缺点多 que-dian-duo '*many shortcomings*' conveys negative sentiment. Longer context generally causes more errors.

8. CONCLUSIONS AND FUTURE WORK

This paper contributes to the research on sentiment classification, domain adaptation and the development of ensembles of complementary classifiers, especially on product reviews written in Chinese. Specifically, we (1) propose a novel ensemble approach (the SELC Model), which successfully integrates a corpus-based model with a lexicon-based approach, (2) present several strategies to overcome the positive classification bias of lexicon-based methods, including the use of a positive/negative ratio control in the iteration process, the use of a general sentiment dictionary to replace a seed word set generated automatically, and the enlargement of negation word list.

Experiments show the effectiveness of the SELC Model. Moreover, the standard deviation of F1-scores on ten domains is reduced, which shows that the SELC Model could be more suitable for domain-independent sentiment classification.

Although our method achieves significant improvement over the previous study, there are still several other avenues that might be explored in future work. First, the use of linguistic knowledge in sentiment classification needs further study. There are many complicated constructions involved in the indirection expression of negative sentiment, and negation word is only one kind of them. For instance, the verb 实现 shi-xian '*achieve*' usually conveys positive sentiment and 避免 bi-mian '*avoid*' negative sentiment, but they are not considered as sentiment words in most sentiment dictionaries. Second, although experiments were conducted only on Chinese corpus in this paper, our model is language-independent theoretically. Therefore, we attempt to apply the model on corpus in other languages.

9. References

[1] Alina Andreevskaia and Sabine Bergler. 2008. When Specialists and Generalists Work Together: Overcoming Domain Dependence in Sentiment Tagging. In *Proceedings*

of the 46th Annual Meeting of the Association for Computational Linguistics, pages 290-298.

[2] Anthony Aue and Michael Gamon. 2005. Customizing sentiment classifiers to new domains: a case study. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, Borovets, BG.

[3] Bo Pang and Lilian Lee. 2004. A sentiment education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics*.

[4] Bo Pang, Lilian Lee, and Shrivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of Conference on Empirical Methods in Natural Language Processing 2002*.

[5] Ethem Alpaydin. 2004. *Introduction to Machine Learning*. The MIT Press, Cambridge, MA.

[6] H. Yu and V. Hatzivassiloglou. 2003. Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. In *Proceedings of Conference on Empirical Methods in Natural Language Processing 2003*.

[7] Hang Cui, Vibhu Mittal, and Mayur Datar. 2006. Comparative experiments on sentiment classification for online product reviews. In *Proceedings of the 21st International Conference on Artificial Intelligence*, Boston, MA.

[8] J.C Na, H. Sui, C. Khoo, S. Chan, & Y. Zhou. 2004. Effectiveness of simple linguistic processing in automatic sentiment classification of product reviews. In *I.C. McIlwaine (Ed.), Knowledge Organization and the Global Information Society: Proceedings of the Eighth International ISKO Conference* (pages 49-54). Wurzburg, Germany: Ergon Verlag.

[9] Jonathon Read. 2005. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL-2005 Student Research Workshop*, Ann Arbor, MI.

[10] Kushal Dave, Steve Lawrence, and David M. Pennock. 2003. Mining the Peanut gallery: opinion extraction and semantic classification of product reviews. In *Proceedings of WWW2003*, Budapest, HU.

[11] Mark Drezde, John Blitzer, and Fernando Pereira. 2007. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, Prague, CZ.

[12] Michael Gamon and Anthony Aue. 2005. Automatic identification of sentiment vocabulary: exploiting low association with known sentiment terms. In *Proceedings of the ACL-05 Workshop on Feature Engineering for Machine Learning in Natural Language Processing*, Ann Arbor, US.

[13] Peter Turney. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics*.

[14] Songbo Tan, Gaowei Wu, Huifeng Tang, and Zueqi Cheng. 2007. A Novel Scheme for Domain-transfer Problem in the context of Sentiment Analysis. In *Proceedings of CIKM 2007*.

[15] Taras Zagibalov and John Carroll. 2008a. Unsupervised Classification of Sentiment and Objectivity in Chinese Text. In *Proceedings of the Third International Joint Conference on Natural Language Processing*. 304– 311.

[16] Taras Zagibalov and John Carroll. 2008b. Automatic Seed Word Selection for Unsupervised Sentiment Classification of Chinese Test. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 107.