Sentiment Classification and Polarity Shifting

Shoushan Li[†‡] **Sophia Yat Mei Lee**[†] **Ying Chen**[†] **Chu-Ren Huang**[†] **Guodong Zhou**[‡]

[†]Department of CBS
The Hong Kong Polytechnic University
{shoushan.li, sophiaym, chenying3176, churenhuang}@gmail.com

[‡] Natural Language Processing Lab
School of Computer Science and Technology
Soochow University
gdzhou@suda.edu.cn

Abstract

Polarity shifting marked by various linguistic structures has been a challenge to automatic sentiment classification. In this paper, we propose a machine learning approach to incorporate polarity shifting information into a document-level sentiment classification system. First, a feature selection method is adopted to automatically generate the training data for a binary classifier on polarity shifting detection of sentences. Then, by using the obtained binary classifier, each document in the original polarity classification training data is split into two partitions, polarity-shifted and polarity-unshifted, which are used to train two base classifiers respectively for further classifier combination. The experimental results across four different domains demonstrate the effectiveness of our approach.

1 Introduction

Sentiment classification is a special task of text classification whose objective is to classify a text according to the sentimental polarities of opinions it contains (Pang et al., 2002), e.g., *favorable* or *unfavorable*, *positive* or *negative*. This task has received considerable interests in the computational linguistic community due to its potential applications.

In the literature, machine learning approaches have dominated the research in sentiment classification and achieved the state-of-the-art performance (e.g., Kennedy and Inkpen, 2006; Pang et al., 2002). In a typical machine learning approach, a document (text) is modeled as a bag-of-words, i.e. a set of content words without any word order or syntactic relation information. In other words, the underlying assumption is that the sentimental orientation of the whole text depends on the sum of the sentimental polarities of content words. Although this assumption is reasonable and has led to initial success, it is linguistically unsound since many function words and constructions can shift the sentimental polarities of a text. For example, in the sentence '*The chair is not comfortable*', the polarity of the word '*comfortable*' is positive while the polarity of the whole sentence is reversed because of the negation word '*not*'. Therefore, the overall sentiment of a document is not necessarily the sum of the content parts (Turney, 2002). This phenomenon is one main reason why machine learning approaches fail under some circumstances.

As a typical case of polarity shifting, negation has been paid close attention and widely studied in the literature (Na et al., 2004; Wilson et al., 2009; Kennedy and Inkpen, 2006). Generally, there are two steps to incorporate negation information into a system: negation detection and negation classification. For negation detection, some negation trigger words, such as '*no*', '*not*', and '*never*', are usually applied to recognize negation phrases or sentences. As for negation classification, one way to import negation information is to directly reverse the polarity of the words which contain negation trigger words as far as term-counting approaches are considered (Kennedy and Inkpen, 2006). An alternative way is to add some negation features (e.g., negation bigrams or negation phrases) into

Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 635–643,
Beijing, August 2010

machine learning approaches (Na et al., 2004). Such approaches have achieved certain success.

There are, however, some shortcomings with current approaches in incorporating negation information. In terms of negation detection, firstly, the negation trigger word dictionary is either manually constructed or relies on existing resources. This leads to certain limitations concerning the quality and coverage of the dictionary. Secondly, it is difficult to adapt negation detection to other languages due to its language dependence nature of negation constructions and words. Thirdly, apart from negation, many other phenomena, e.g., contrast transition with trigger words like '*but*', '*however*', and '*nevertheless*', can shift the sentimental polarity of a phrase or sentence. Therefore, considering negation alone is inadequate to deal with the polarity shifting problem, especially for document-level sentiment classification.

In terms of negation classification, although it is easy for term-counting approaches to integrate negation information, they rarely outperform a machine learning baseline (Kennedy and Inkpen, 2006). Even for machine learning approaches, although negation information is sometimes effective for local cases (e.g., *not good*), it fails on long-distance cases (e.g., *I don't think it is good*).

In this paper, we first propose a feature selection method to automatically generate a large scale polarity shifting training data for polarity shifting detection of sentences. Then, a classifier combination method is presented for incorporating polarity shifting information. Compared with previous ones, our approach highlights the following advantages: First of all, we apply a binary classifier to detect polarity shifting rather than merely relying on trigger words or phrases. This enables our approach to handle different kinds of polarity shifting phenomena. More importantly, a feature selection method is presented to automatically generate the labeled training data for polarity shifting detection of sentences.

The remainder of this paper is organized as follows. Section 2 introduces the related work of sentiment classification. Section 3 presents our approach in details. Experimental results are presented and analyzed in Section 4. Finally, Section 5 draws the conclusion and outlines the future work.

2 Related Work

Generally, sentiment classification can be performed at four different levels: word level (Wiebe, 2000), phrase level (Wilson et al., 2009), sentence level (Kim and Hovy, 2004; Liu et al., 2005), and document level (Turney, 2002; Pang et al., 2002; Pang and Lee, 2004; Riloff et al., 2006). This paper focuses on document-level sentiment classification.

In the literature, there are mainly two kinds of approaches on document-level sentiment classification: term-counting approaches (lexicon-based) and machine learning approaches (corpus-based). Term-counting approaches usually involve deriving a sentiment measure by calculating the total number of negative and positive terms (Turney, 2002; Kim and Hovy, 2004; Kennedy and Inkpen, 2006). Machine learning approaches recast the sentiment classification problem as a statistical classification task (Pang and Lee, 2004). Compared to term-counting approaches, machine learning approaches usually achieve much better performance (Pang et al., 2002; Kennedy and Inkpen, 2006), and have been adopted to more complicated scenarios, such as domain adaptation (Blitzer et al., 2007), multi-domain learning (Li and Zong, 2008) and semi-supervised learning (Wan, 2009; Dasgupta and Ng, 2009) for sentiment classification.

Polarity shifting plays a crucial role in phrase-level, sentence-level, and document-level sentiment classification. However, most of previous studies merely focus on negation shifting (polarity shifting caused by the negation structure). As one pioneer research on sentiment classification, Pang et al. (2002) propose a machine learning approach to tackle negation shifting by adding the tag 'not' to every word between a negation trigger word/phrase (e.g., *not*, *isn't*, *didn't*, etc.) and the first punctuation mark following the negation trigger word/phrase. To their disappointment, considering negation shifting has a negligible effect and even slightly harms the overall performance. Kennedy and Inkpen (2006) explore negation shifting by incorporating negation bigrams as additional features into machine learning approaches. The

experimental results show that considering sentiment shifting greatly improves the performance of term-counting approaches but only slightly improves the performance of machine learning approaches. Other studies such as Na et al. (2004), Ding et al. (2008), and Wilson et al. (2009) also explore negation shifting and achieve some improvements[1]. Nonetheless, as far as machine learning approaches are concerned, the improvement is rather insignificant (normally less than 1%). More recently, Ikeda et al. (2008) first propose a machine learning approach to detect polarity shifting for sentence-level sentiment classification, based on a manually-constructed dictionary containing thousands of positive and negative sentimental words, and then adopt a term-counting approach to incorporate polarity shifting information.

3 Sentiment Classification with Polarity Shifting Detection

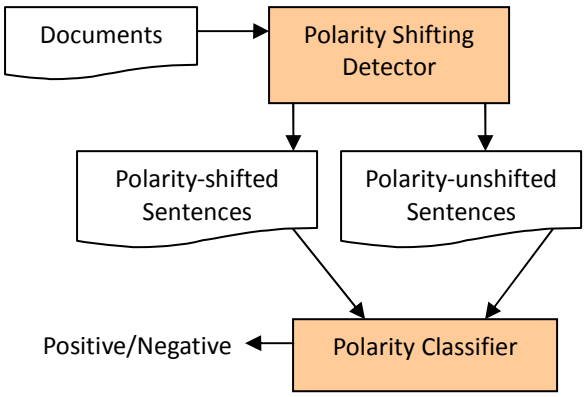

Figure 1: General framework of our approach

The motivation of our approach is to improve the performance of sentiment classification by robust treatment of sentiment polarity shifting between sentences. With the help of a binary classifier, the sentences in a document are divided into two parts: sentences which contain polarity shifting structures and sentences without any polarity shifting structure. Figure 1 illustrates the general framework of our approach. Note that this framework is a general one, that is, different polarity shifting detection methods can be applied to differentiate polarity-shifted sentences from those polarity-unshifted sentences and different polarity classification methods can be adopted to incorporate sentiment shifting information. For clarification, the training data used for polarity shifting detection and polarity classification are referred to as the polarity shifting training data and the polarity classification training data, respectively.

3.1 Polarity Shifting Detection

In this paper, polarity shifting means that the polarity of a sentence is different from the polarity expressed by the sum of the content words in the sentence. For example, in the sentence "*I am not disappointed*", the negation structure makes the polarity of the word '*disappointed*' different from that of the whole sentence (*negative* vs. *positive*). Apart from the negation structure, many other linguistic structures allow polarity shifting, such as contrast transition, modals, and pre-suppositional items (Polanyi and Zaenen, 2006). We refer these structures as polarity shifting structures.

One of the great challenges in building a polarity shifting detector lies on the lack of relevant training data since manually creating a large scale corpus of polarity shifting sentences is time-consuming and labor-intensive. Ikeda et al. (2008) propose an automatic way for collecting the polarity shifting training data based on a manually-constructed large-scale dictionary. Instead, we adopt a feature selection method to build a large scale training corpus of polarity shifting sentences, given only the already available document-level polarity classification training data. With the help of the feature selection method, the top-ranked word features with strong sentimental polarity orientation, e.g., '*great*', '*love*', '*worst*' are first chosen as the polarity trigger words. Then, those sentences with the top-ranked polarity trigger words in both categories of positive and negative documents are selected. Finally, those candidate sentences taking opposite-polarity compared to the containing trigger word are deemed as polarity-shifted.

The basic idea of automatically generating the polarity shifting training data is based on the assumption that the real polarity of a word or phrase is decided by the major polarity category where the word or phrase appears more often. As a result, the sentences in the

[1] Note that Ding et al. (2006) also consider *but*-clause, another important structure for sentiment shifting. Wilson et al. (2009) use conjunctive and dependency relations among polarity words.

frequently-occurring category would be seen as polarity-unshifted while the sentences in the infrequently-occurring category would be seen as polarity-shifted.

In the literature, various feature selection methods, such as Mutual Information (MI), Information Gain (IG) and Bi-Normal Separation (BNS) (Yang and Pedersen, 1997; Forman 2003), have been employed to cope with the problem of the high-dimensional feature space which is normal in sentiment classification.

In this paper, we employ the theoretical framework, proposed by Li et al. (2009), including two basic measurements, i.e. *frequency measurement* and *ratio measurement*, where the first measures, the document frequency of a term in one category, and the second measures, the ratio between the document frequency in one category and other categories. In particular, a novel method called Weighed Frequency and Odds (WFO) is proposed to incorporate both basic measurements:

$$WFO(t,c_i) = P(t \mid c_i)^{\lambda}\{\max(0, \log\frac{P(t \mid c_i)}{P(t \mid \overline{c_i})})\}^{1-\lambda}$$

where $P(t \mid c_i)$ denotes the probability that a document *x* contains the term *t* with the condition that *x* belongs to category c_i; $P(t \mid \overline{c_i})$ denotes the probability that a document *x* contains the term *t* with the condition that *x* does not belong to category c_i. The left part of the formula $P(t \mid c_i)$ implies the first basic measurement and the right part $\log(P(t \mid c_i)/P(t \mid \overline{c_i}))$ implies the second one. The parameter λ $(0 \le \lambda \le 1)$ is thus to tune the weight between the two basic measurements. Especially, when λ equals 0, the WFO method fades to the MI method which fully prefers the second basic measurement.

Figure 2 illustrates our algorithm for automatically generating the polarity shifting training data where c_1 and c_2 denote the two sentimental orientation categories, i.e. negative and positive. *Step A* segments a document into sentences with punctuations. Besides, two special words, '*but*' and '*and*', are used to further segment some contrast transition structures and compound sentences. *Step B* employs the WFO method to rank all features including the words. *Step D* extracts those polarity-shifted and polarity-unshifted sentences containing t_{top-i} where $N_{\max}$ denotes the upper-limit number of sentences in each category of the polarity shifting training data and *#(x)* denotes the total number of the elements in *x*. Apart from that, the first word in the following sentence is also included to capture a common kind of long-distance polarity shifting structure: contrast transition. Thus, important trigger words like '*however*' and '*but*' may be considered. Finally, *Step E* guarantees the balance between the two categories of the polarity shifting training data.

Given the polarity shifting training data, we apply SVM classification algorithm to train a polarity-shifting detector with word unigram features.

Input:

The polarity classification training data: the negative sentimental document set D_{c_1} and the positive sentimental document set D_{c_2}.

Output:

The polarity shifting training data: the polarity-unshifted sentence set $S_{unshift}$ and the polarity-shifted sentence set S_{shift}.

Procedure:

A. *Segment documents* D_{c_1} *and* D_{c_2} *to single sentences* S_{c_1} *and* S_{c_2}.

B. *Apply feature selection on the polarity classification training data and get the ranked features,* $(t_{top-1},...,t_{top-i},...,t_{top-N})$

C. $S_{shift} = \{\}$, $S_{unshift} = \{\}$

D. *For* t_{top-i} *in* $(t_{top-1},...,t_{top-i},...,t_{top-N})$ *:*

D1) if $\#(S_{shift}) > N_{\max}$ *: break*

D2) Collect all sentences S_{top-i,c_1} *and* S_{top-i,c_2} *which contain* t_{top-i} *from* S_{c_1} *and* S_{c_2} *respectively*

D3) if $\#(S_{top-i,c_1}) > \#(S_{top-i,c_2})$ *:*

put S_{top-i,c_2} *into* S_{shift}

put S_{top-i,c_1} *into* $S_{unshift}$

else:

put S_{top-i,c_1} *into* S_{shift}

put S_{top-i,c_2} *into* $S_{unshift}$

E. *Randomly select* $N_{\max}$ *sentences from* $S_{unshift}$ *as the output of* $S_{unshift}$

Figure 2: The algorithm for automatically generating the polarity shifting training data

3.2 Polarity Classification with Classifier Combination

After polarity shifting detection, each document in the polarity classification training data is divided into two parts, one containing polarity-shifted sentences and the other containing polarity-unshifted sentences, which are used to form the polarity-shifted training data and the polarity-unshifted training data. In this way, two different polarity classifiers, f_1 and f_2, can be trained on the polarity-shifted training data and the polarity-unshifted training data respectively. Along with classifier f_3, trained on all original polarity classification training data, we now have three base classifiers in hand for possible classifier combination via a multiple classifier system.

The key issue in constructing a multiple classifier system (MCS) is to find a suitable way to combine the outputs of the base classifiers. In MCS literature, various methods are available for combining the outputs, such as fixed rules including the voting rule, the product rule and the sum rule (Kittler et al., 1998) and trained rules including the weighted sum rule (Fumera and Roli, 2005) and the meta-learning approaches (Vilalta and Drissi, 2002). In this study, we employ the product rule, a popular fixed rule, and stacking (Džeroski and Ženko, 2004), a well-known trained rule, to combine the outputs.

Formally, each base classifier provides some kind of confidence measurements, e.g., posterior probabilities of the test sample belonging to each class. Formally, each base classifier f_l $(l=1,2,3)$ assigns a test sample (denoted as x_l) a posterior probability vector $\vec{P}(x_l)$:

$$\vec{P}(x_l) = (p(c_1 \mid x_l), p(c_2 \mid x_l))^t$$

where $p(c_1 \mid x_l)$ denotes the probability that the l-th base classifier considers the sample belonging c_1.

The product rule combines the base classifiers by multiplying the posterior possibilities and using the multiplied possibility for decision, i.e.

$$assign \quad y \to c_j \ when \ j = \arg\max_i \prod_{l=1}^{3} p(c_i \mid x_l)$$

Stacking belongs to well-known meta-learning (Vilalta and Drissi, 2002). The key idea behind meta-learning is to train a meta-classifier with input attributes that are the outputs of the base classifiers. Hence, meta-learning usually needs some development data for generating the meta-training data. Let x' denote a feature vector of a sample from the development data. The output of the l-th base classifier f_l on this sample is the probability distribution over the category set $\{c_1, c_2\}$, i.e.

$$\vec{P}(x'_l) = (p(c_1 \mid x'_l), p_l(c_2 \mid x'_l))$$

A meta-classifier can be trained using the development data with the meta-level feature vector $x^{meta} \in R^{2\times 3}$

$$x^{meta} = (\vec{P}(x'_{l=1}), \vec{P}(x'_{l=2}), \vec{P}(x'_{l=3}))$$

Stacking is a specific meta-learning rule, in which a leave-one-out or a cross-validation procedure on the training data is applied to generate the meta-training data instead of using extra development data. In our experiments, we perform stacking with 10-fold cross-validation to generate the meta-training data.

4 Experimentation

4.1 Experimental Setting

The experiments are carried out on product reviews from four domains: books, DVDs, electronics, and kitchen appliances (Blitzer et al., 2007)[2]. Each domain contains 1000 positive and 1000 negative reviews.

For sentiment classification, all classifiers including the polarity shifting detector, three base classifiers and the meta-classifier in stacking are trained by SVM using the SVM-light tool [3] with Logistic Regression method for probability measuring (Platt, 1999).

In all the experiments, each dataset is randomly and evenly split into two subsets: 50% documents as the training data and the remaining 50% as the test data. The features include word unigrams and bigrams with Boolean weights.

4.2 Experimental Results on Polarity Shifting Data

To better understand the polarity shifting phenomena in document-level sentiment classification, we randomly investigate 200

[2] This data set is collected by Blitzer et al. (2007): http://www.seas.upenn.edu/~mdredze/datasets/sentiment/

[3] It is available at: http://svmlight.joachims.org/

polarity-shifted sentences, together with their contexts (i.e. the sentences before and after it), automatically generated by the WFO ($\lambda = 0$) feature selection method. We find that nearly half of the automatically generated polarity-shifted sentences are actually polarity-unshifted sentences or difficult to decide. That is to say, the polarity shifting training data is noisy to some extent. One main reason is that some automatically selected trigger words do not really contain sentiment information, e.g., '*hear*', '*information*' etc. Another reason is that some reversed opinion is given in a review without any explicit polarity shifting structures.

To gain more insights, we manually checked 100 sentences which are explicitly polarity-shifted and can also be judged by human according to their contexts. Table 1 presents some typical structures causing polarity shifting. It shows that the most common polarity shifting type is Explicit Negation (37%), usually expressed by trigger words such as '*not*', '*no*', or '*without*', e.g., in the sentence '*I am not happy with this flashcard at all*'. Another common type of polarity shifting is Contrast Transition (20%), expressed by trigger words such as '*however*', e.g., in the sentence '*It is large and stylish, however, I cannot recommend it because of the lid*'. Other less common yet productive polarity shifting types include Exception and Until. Exception structure is usually expressed by the trigger phrase '*the only*' to indicate the one and only advantage of the product, e.g., in the sentence '*The only thing that I like about it is that bamboo is a renewable resource*'. Until structure is often expressed by the trigger word '*until*' to show the reversed polarity, e.g. in the sentence '*This unit was a great addition until the probe went bad after only a few months*'.

Polarity Shifting Structures	Trigger Words/Phrases	Distribution (%)
Explicit Negation	*not, no, without*	37
Contrast Transition	*but, however, unfortunately*	20
Implicit Negation	*avoid, hardly,*	7
False Impression	*look, seem*	6
Likelihood	*probably, perhaps*	5
Counter-factual	*should, would*	5
Exception	*the only*	5
Until	*until*	3

Table 1: Statistics on various polarity shifting structures

4.3 Experimental Results on Polarity Classification

For comparison, several classifiers with different classification methods are developed.

1) Baseline classifier, which applies SVM with all unigrams and bigrams. Note that it also serves as a base classifier in the following combined classifiers.

2) Base classifier 1, a base classifier for the classifier combination method. It works on the polarity-unshifted data.

3) Base classifier 2, another base classifier for the classifier combination method. It works on the polarity-shifted data.

4) Negation classifier, which applies SVM with all unigrams and bigrams plus negation bigrams. It is a natural extension of the baseline classifier with the consideration of negation bigrams. In this study, the negation bigrams are collected using some negation trigger words, such as '*not*' and '*never*'. If a negation trigger word is found in a sentence, each word in the sentence is attached with the word '*_not*' to form a negation bigram.

5) Product classifier, which combines the baseline classifier, the base classifier 1 and the base classifier 2 using the product rule.

6) Stacking classifier, a combined classifier similar to the **Product classifier**. It uses the stacking classifier combination method instead of the product rule.

Please note that we do not compare our approach with the one as proposed in Ikeda et al. (2008) due to the absence of a manually-collected sentiment dictionary. Besides, it is well known that a combination strategy itself is capable of improving the classification performance. To justify whether the improvement is due to the combination strategy or our polarity shifting detection or both, we first randomly split the training data into two portions and train two base classifiers on each portion, then apply the stacking method to combine them along with the baseline classifier. The corresponding results are shown as 'Random+Stacking' in Table 2. Finally, in our experiments, *t*-test is performed to evaluate the significance of the performance improvement between two systems employing different methods (Yang and Liu, 1999).

Domain	Baseline	Base Classifier 1	Base Classifier 2	Negation Classifier	Random + Stacking	Shifting + Product	Shifting + Stacking
Book	0.755	0.756	0.670	0.759	0.764	0.772	**0.785**
DVD	0.750	0.743	0.667	0.748	0.759	0.768	**0.770**
Electronic	0.779	0.786	0.711	0.785	0.789	0.820	**0.830**
Kitchen	0.818	0.814	0.683	0.826	0.835	0.840	**0.849**

Table 2: Performance comparison of different classifiers with equally-splitting between training and test data

Performance comparison of different classifiers

Table 2 shows the accuracy results of different methods using 2000 polarity shifted sentences and 2000 polarity-unshifted sentences to train the polarity shifting detector (N_{max}=2000). Compared to the baseline classifier, it shows that: 1) The base classifier 1, which only uses the polarity-unshifted sentences as the training data, achieves similar performance. 2) The base classifier 2 achieves much lower performance due to much fewer sentences involved. 3) Including negation bigrams usually allows insignificant improvements (*p-value*>0.1), which is consistent with most of previous works (Pang et al., 2002; Kennedy and Inkpen, 2006). 4) Both the product and stacking classifiers with polarity shifting detection significantly improve the performance (*p-value*<0.05). Compared to the product rule, the stacking classifier is preferable, probably due to the performance unbalance among the individual classifiers, e.g., the performance of the base classifier 2 is much lower than the other two. Although stacking with two randomly generated base classifiers, i.e. "Random + Stacking", also consistently outperforms the baseline classifier, the improvements are much lower than what has been achieved by our approach. This suggests that both the classifier combination strategy and polarity shifting detection contribute to the overall performance improvement.

Effect of WFO feature selection method

Figure 3 presents the accuracy curve of the stacking classifier when using different Lambda (λ) values in the WFO feature selection method. It shows that those feature selection methods which prefer frequency information, e.g., MI and BNS, are better in automatically generating the polarity shifting training data. This is reasonable since high frequency terms, e.g., '*is*', '*it*', '*a*', etc., tend to obey our assumption that the real polarity of one top term should belong to the polarity category where the term appears frequently.

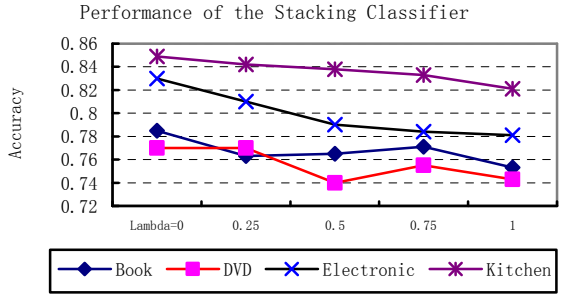

Figure 3: Performance of the stacking classifier using WFO with different Lambda (λ) values

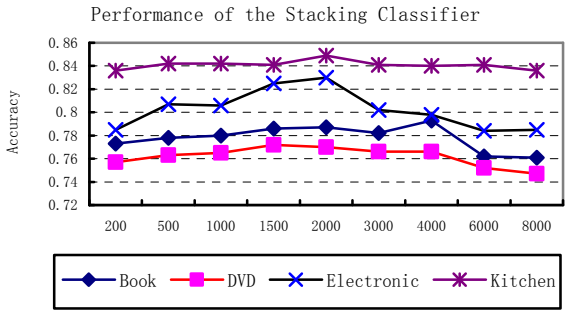

Figure 4: Performance of the stacking classifier over different sizes of the polarity shifting training data (with N_{max} sentences in each category)

Effect of a classifier over different sizes of the polarity shifting training data

Another factor which might influence the overall performance is the size of the polarity shifting training data. Figure 4 presents the overall performance on different numbers of the polarity shifting sentences when using the stacking classifier. It shows that 1000 to 4000 sentences are enough for the performance improvement. When the number is too large, the noisy training data may harm polarity shifting detection. When the number is too small, it is not enough for the automatically generated polarity shifting training data to capture various polarity shifting structures.

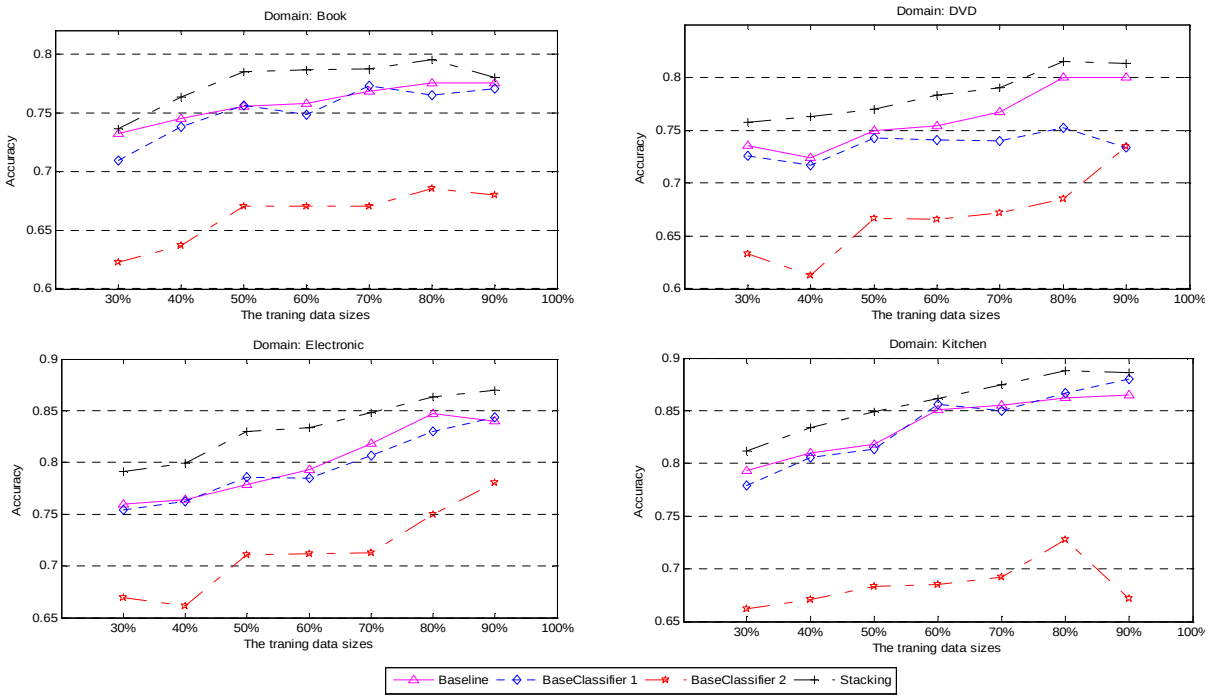

Figure 5: Performance of different classifiers over different sizes of the polarity classification training data

Effect of different classifiers over different sizes of the polarity classification training data

Figure 5 shows the classification results of different classifiers with varying sizes of the polarity classification training data. It shows that our approach is able to improve the overall performance robustly. We also notice the big difference between the performance of the baseline classifier and that of the base classifier 1 when using 30% training data in Book domain and 90% training data in DVD domain. Detailed exploration of the polarity shifting sentences in the training data shows that this difference is mainly attributed to the poor performance of the polarity shifting detector. Even so, the stacking classifier guarantees no worse performance than the baseline classifier.

5 Conclusion and Future Work

In this paper, we propose a novel approach to incorporate polarity shifting information into document-level sentiment classification. In our approach, we first propose a machine-learning-based classifier to detect polarity shifting and then apply two classifier combination methods to perform polarity classification. Particularly, the polarity shifting training data is automatically generated through a feature selection method. As shown in our experimental results, our approach is able to consistently improve the overall performance across different domains and training data sizes, although the automatically generated polarity shifting training data is prone to noise. Furthermore, we conclude that those feature selection methods, which prefer frequency information, e.g., MI and BNS, are good choices for generating the polarity shifting training data.

In our future work, we will explore better ways in generating less-noisy polarity shifting training data. In addition, since our approach is language-independent, it is readily applicable to sentiment classification tasks in other languages.

For availability of the automatically generated polarity shifting training data, please contact the first author (for research purpose only).

Acknowledgments

This research work has been partially supported by Start-up Grant for Newly Appointed Professors, No. 1-BBZM in the Hong Kong Polytechnic University and two NSFC grants, No. 60873150 and No. 90920004. We also thank the three anonymous reviewers for their helpful comments.

References

Blitzer J., M. Dredze, and F. Pereira. 2007. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. In *Proceedings of ACL-07*.

Dasgupta S. and V. Ng. 2009. Mine the Easy and Classify the Hard: Experiments with Automatic Sentiment Classification. In *Proceedings of ACL-IJCNLP-09*.

Ding X., B. Liu, and P. Yu. 2008. A Holistic Lexicon-based Approach to Opinion Mining. In *Proceedings of the International Conference on Web Search and Web Data Mining, WSDM-08*.

Džeroski S. and B. Ženko. 2004. Is Combining Classifiers with Stacking Better than Selecting the Best One? *Machine Learning*, vol.54(3), pp.255-273, 2004.

Forman G. 2003. An Extensive Empirical Study of Feature Selection Metrics for Text Classification. *The Journal of Machine Learning Research*, 3(1), pp.1289-1305.

Fumera G. and F. Roli. 2005. A Theoretical and Experimental Analysis of Linear Combiners for Multiple Classifier Systems. *IEEE Trans. PAMI*, vol.27, pp.942–956, 2005

Ikeda D., H. Takamura, L. Ratinov, and M. Okumura. 2008. Learning to Shift the Polarity of Words for Sentiment Classification. In *Proceedings of IJCNLP-08*.

Kennedy, A. and D. Inkpen. 2006. Sentiment Classification of Movie Reviews using Contextual Valence Shifters. *Computational Intelligence*, vol.22(2), pp.110-125, 2006.

Kim S. and E. Hovy. 2004. Determining the Sentiment of Opinions. In *Proceedings of COLING-04*.

Kittler J., M. Hatef, R. Duin, and J. Matas. 1998. On Combining Classifiers. *IEEE Trans. PAMI*, vol.20, pp.226-239, 1998

Li S., R. Xia, C. Zong, and C. Huang. 2009. A Framework of Feature Selection Methods for Text Categorization. In *Proceedings of ACL-IJCNLP-09*.

Li S. and C. Zong. 2008. Multi-domain Sentiment Classification. In *Proceedings of ACL-08: HLT*, short paper.

Liu B., M. Hu, and J. Cheng. 2005. Opinion Observer: Analyzing and Comparing Opinions on the Web. In *Proceedings of WWW-05*.

Na J., H. Sui, C. Khoo, S. Chan, and Y. Zhou. 2004. Effectiveness of Simple Linguistic Processing in Automatic Sentiment Classification of Product Reviews. In *Conference of the International Society for Knowledge Organization (ISKO-04)*.

Pang B. and L. Lee. 2004. A Sentimental Education: Sentiment Analysis using Subjectivity Summarization based on Minimum Cuts. In *Proceedings of ACL-04*.

Pang B., L. Lee, and S. Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proceedings of EMNLP-02*.

Platt J. 1999. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In: A. Smola, P. Bartlett, B. Schoelkopf and D. Schuurmans (Eds.): *Advances in Large Margin Classiers*. MIT Press, Cambridge, 61-74.

Polanyi L. and A. Zaenen. 2006. Contextual Valence Shifters. *Computing attitude and affect in text: Theory and application*. Springer Verlag.

Riloff E., S. Patwardhan, and J. Wiebe. 2006. Feature Subsumption for Opinion Analysis. In *Proceedings of EMNLP-06*.

Turney P. 2002. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proceedings of ACL-02*.

Vilalta R. and Y. Drissi. 2002. A Perspective View and Survey of Meta-learning. *Artificial Intelligence Review*, 18(2), pp. 77–95.

Wan X. 2009. Co-Training for Cross-Lingual Sentiment Classification. In *Proceedings of ACL-IJCNLP-09*.

Wiebe J. 2000. Learning Subjective Adjectives from Corpora. In *Proceedings of AAAI-2000*.

Wilson T., J. Wiebe, and P. Hoffmann. 2009. Recognizing Contextual Polarity: An Exploration of Features for Phrase-Level Sentiment Analysis. *Computational Linguistics*, vol.35(3), pp.399-433, 2009.

Yang Y. and X. Liu, X. 1999. A Re-Examination of Text Categorization methods. In *Proceedings of SIGIR-99*.

Yang Y. and J. Pedersen. 1997. A Comparative Study on Feature Selection in Text Categorization. In *Proceedings of ICML-97*.