Information Processing and Management 52 (2016) 36–45

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Polarity shift detection, elimination and ensemble: A three-stage model for document-level sentiment analysis

Rui Xia[a,d], Feng Xu[b], Jianfei Yu[a], Yong Qi[a,b], Erik Cambria[c,*]

[a] School of Computer Science and Engineering, Nanjing University of Science and Technology, China
[b] School of Economics and Management, Nanjing University of Science and Technology, China
[c] School of Computer Engineering, Nanyang Technological University, Singapore
[d] State Key Laboratory for Novel Software Technology, Nanjing University, China

ARTICLE INFO

Article history:
Received 26 June 2014
Revised 31 March 2015
Accepted 10 April 2015
Available online 11 November 2015

Keywords:
Sentiment analysis
Sentiment classification
Polarity shift

ABSTRACT

The polarity shift problem is a major factor that affects classification performance of machine-learning-based sentiment analysis systems. In this paper, we propose a three-stage cascade model to address the polarity shift problem in the context of document-level sentiment classification. We first split each document into a set of subsentences and build a hybrid model that employs rules and statistical methods to detect explicit and implicit polarity shifts, respectively. Secondly, we propose a polarity shift elimination method, to remove polarity shift in negations. Finally, we train base classifiers on training subsets divided by different types of polarity shifts, and use a weighted combination of the component classifiers for sentiment classification. The results on a range of experiments illustrate that our approach significantly outperforms several alternative methods for polarity shift detection and elimination.

1. Introduction

The volume of user-generated text on the Web in the form of reviews, blogs, and social networks has grown dramatically in recent years. This was mirrored by an increasing interest, from both the academic and the business world, in the field of sentiment analysis, which aims to automatically extract sentiment from natural language text and can be broadly categorized into knowledge-based (Cambria et al., 2013) or statistics-based (Cambria et al., 2013b). While the former leverages on the use of ontologies (Gangemi, Presutti, & Reforgiato, 2014) and semantic networks (Cambria, Hussain, Havasi, & Eckl, 2010a) to infer sentiments from text in an unsupervised way, the latter focuses on machine learning (Poria, Gelbukh, Cambria, Hussain, & Huang, 2014; Xia et al., 2015) and clustering techniques (Cambria, Mazzocco, Hussain, & Eckl, 2011) to detect polarity in a supervised way.

In standard practice, sentiment analysis is considered as a special case of text classification, where a review text is normally represented by a bag-of-words (BOW) model. Then, statistical machine learning algorithms, such as Naïve Bayes, maximum entropy classifier, and support vector machine (SVM) are used for classification. However, the BOW model disrupts word order, breaks the syntactic structures and discards some semantic information of the text. Therefore, it brings about some fundamental deficiencies including the polarity shift problem. Polarity shift refers to a linguistic phenomenon in which the polarity of sentiment can be reversed (i.e., positive to negative or vice versa) by some special linguistic structures

* Corresponding author.
E-mail address: cambria@ntu.edu.sg (E. Cambria).

http://dx.doi.org/10.1016/j.ipm.2015.04.003
0306-4573/

called polarity shifters, e.g., negation ("*I don't like this movie*") and contrast ("*Fairly good, but not my style*"). Obviously, in the BOW model, it is hard to capture the sentiment reversion caused by polarity shifters, because two sentiment-opposite texts (e.g., "*I don't like this movie*" and "*I like this movie*") are regarded to be very similar in the BOW representation.

Several approaches have been proposed in the literature to address the polarity shift problem (Das, 2001; Choi, 2008; Kennedy & Inkpen, 2006; Li, Chen, & Zhou, 2010; Li & Zong, 2009; Na, Khoo, & Zhou, 2004; Polanyi, 2004; Wilson & Hoffmann, 2005). However, most of them focused on either modeling polarity shift in phrase/subsentence-level sentiment classification, or encoding polarity shift in rule-based term-counting methods. Even there were few of them dealing with polarity shift by using machine learning methods for document-level sentiment classification, their performances were not satisfactory, e.g., the improvements were less than 2% after considering polarity shift in (Li et al., 2010).

In this work, we propose a three-stage model, namely Polarity Shift Detection, Elimination and Ensemble (PSDEE), to address polarity shift for document-level sentiment classification. Firstly, we propose a hybrid polarity shift detection approach, which employs a rule-based method to detect some polarity shifts such as explicit negations and contrasts, and a statistical method to detect some implicit polarity shifts such as sentiment inconsistencies. Secondly, we propose a novel polarity shift elimination algorithm to eliminate polarity shifts in negations. For example, the review "*this movie is not interesting*" is reversed to "*this movie is boring*". It can make the BOW representation more feasible due to the elimination of negations. Finally, we separate the training and test data into four component subsets, i.e., negation subset, contrast subset, sentiment-inconsistency set as well as polarity-unshifted subset, and train the base classifiers based on each of the component subset. A weighted ensemble of four component predictions are finally used in testing, with the motivation to distinguish texts with different types of polarity shifts such that the polarity-unshifted part will have a higher weight, while the polarity-shifted part will have a lower weight in sentiment prediction. We systematically evaluate our PSDEE model by conducting experiments on four sentiment datasets, three kinds of classification algorithms and two types of features. The experimental results prove the effectiveness of our PSDEE model across different settings.

The rest of the paper is organized as follows: Section 2 presents the motivation; in Section 3, we introduce our PSDEE model in detail by discussing (a) the hybrid polarity shift detection method, (b) the negation elimination approach, and (c) the polarity-shift-based ensemble model; experimental results are reported and analyzed in Section 4; we review related work in Section 5; finally, Section 6 draws the conclusions.

2. Motivation

2.1. How to detect different types of polarity shifts?

Polarity shifters, also called "valence shifter" in (Polanyi & Zaenen, 2004) and "sentiment shifter" in (Liu, 2012) are words and phrases that can change sentiment orientations of texts. Polarity shift is a complex linguistic structure that may include explicit negations, contrasts, intensifiers, diminishers, irrealis, etc. (Taboada, Brooke, Tofiloski, Voll, & Stede, 2011). Li et al. (2010) have conducted a statistic on the distribution of different types of polarity shift, and reported that explicit negations and contrasts covers more than 60% polarity shift structures.

Negation is the most common type of polarity shifts. For example, in the review:

Review 1 (Explicit Negation). *"I don't like this movie"*, the negator "*don't*" shifts the polarity of the sentiment word "*like*". It usually has explicit hints (i.e., negators) in negation. Therefore, we can capture the explicit negation by using some rule-based methods based on the presence of some pre-defined negators.

Contrast is another important class of polarity shifts. For example in the review:

Review 2 (Explicit Contrast). *"Fairly good acting, but overall a disappointing movie"*, the contrast indicator "but" shifts the sentiment polarity of the previous phrase "Fairly good acting". Similar as explicit negations, we may also use rule-based method to detect the explicit contrasts according to some pre-defined contrast indicators.

While some polarity shift structures such as explicit negations and contrasts are relative easy to detect, there still exists a large part of implicit polarity shifts that are very hard to detect based on simple rule-based methods. For example in the review:

Review 3 (Sentiment Inconsistency). *"I don't like this movie. Great actor, awful scenario*", the first phrase "*I don't like this movie*" expresses a negative sentiment toward the whole film, the second phrase "*great actor*" shows a positive sentiment toward acting, and the third phrase "*awful scenario*" expresses the negative sentiment toward the aspect of scenario. In this case, people hold an opposite opinion toward one subordinate aspect, which is opposite to the sentiment of the whole review. We call this type of polarity shift "sentiment inconsistency". Pang, Lee, and Vaithyanathan (2002) referred to this problem as "thwarted expectation", which is synonymous to inconsistent or mixed sentiment patterns in the review text. This phenomenon is especially common in long review texts, where people might have different opinions toward different aspects of one product. But in sentiment inconsistencies, the opinion is inconsistent to that in its neighbors, and is always contrary to the sentiment expressed on the product overall. In this case, there are not explicit hints for polarity shift detection. Nevertheless, we could use a statistical method to detect them.

2.2. How to handle different types of polarity shifts in sentiment classification?

Different types of polarity shifts should be handled in different ways. In most situations, negators can only shift the sentiment polarity within the phrase. If negation (i.e., "*don't like*") is not handled in the text representation, the sentiment word in the scope of negation (i.e., "*like*") will play an opposite role in sentiment classification. In this work, we propose a novel method, called polarity shift elimination, to remove the negators and the same time reverse the sentiment word to their antonyms. For example in Review 1, the text "*I don't like this movie*" is converted to a new one "*I dislike this movie*", where "*dislike*" is the antonymous word of "*like*" according to an antonym dictionary. Negations in the text are removed by polarity shift elimination. Hence, the learning and prediction errors caused by negations in machine learning algorithms will be corrected.

Different from negation, contrast can shift the polarity of its neighboring sentence or subsentence. For example, in Review 2, "*but*" shifts the polarity of its previous subsentence "*fairly good acting*". If contrast polarity shift is not considered, sentiment word in the shifted subsentence (e.g., "*good*") will also play an opposite role in sentiment classification. In this case, we should decrease the impact of the shifted part ("*fairly good acting*"), and increase the impact of unshifted part ("*overall a disappointing movie*") in sentiment classification.

As for sentiment inconsistency, its main cause is that people have different opinions toward different aspects of the product. Hence, the sentiment orientations toward some the less-important aspects may be opposite to the overall sentiment orientation. Therefore, the impact of the sentiment inconsistent part should be reduced, while the impact of the sentiment consistent part should be increased. From this point of view, sentiment inconsistency has some similarities to contrasts, and it can be viewed as a type of implicit contrast. The impact of both the explicit contrasts and sentiment inconsistency (e.g., implicit contrast) should be weakened in sentiment classification.

To meet abovementioned needs, we propose an ensemble model, to leverage the impact of negations, contrasts, sentiment inconsistency and polarity-unshifted text, for building a polarity shift aware sentiment classifier. An effective ensemble is supposed to assign a large weight to the base classifier trained on the polarity-upshifted text, a moderate weight to the negation part, and a small weight to the contrast and sentiment inconsistency part.

3. The PSDEE approach

This work presents a three-stage cascade model for document-level sentiment classification. The three stages are (1) hybrid polarity shift detection, (2) polarity shift elimination in negations, and (3) polarity shift based ensemble model. Fig. 1 gives an illustration of the PSDEE approach.

3.1. Hybrid polarity shift detection

In the first stage, we propose a hybrid model to detect different types of polarity shifts summarized in Section 2.1. Specifically, we employ a rule-based method to detect explicit negations and contrasts, and a statistical method to detect the implicit sentiment inconsistency. Fig. 2 presents the pseudo-code of the polarity shift detection algorithm.

3.1.1. Rule-based polarity detection for negations and explicit contrasts

As we have mentioned, a myriad of polarity shifts such as explicit negations and explicit contrasts have obvious hints. We summarize a set of pre-defined negators and disjunctive conjunctions, use them as hints of the explicit negations and contrasts, and subsequently propose a rule-based method for polarity shift detection.

Let $\mathcal{N} = \{n_1, n_2, \ldots n_t\}$ denotes the set of negation indicators (i.e., negators), $\mathcal{C} = \{c_1, c_2, \ldots c_t\}$ denotes the set of contrast indicators (i.e., disjunctive conjunctions).[1] Suppose a document d is composed of m subsentences $d = (s_1, s_2, \ldots s_m)$, where each subsentence s_i is represented by a list of words contained in the sentence $s_i = (w_{i1}, w_{i2}, \ldots w_{i|s_i|})$. Define d_{negation} and d_{contrast} as subsets of d that contain negations and contrasts, respectively. Line 3–10 in Fig. 2 shows the rule-based methods for detecting negations and explicit contrasts. Specifically, we put the subsentence s_i that contains a negation indicator into d_{negation}. For a subsentence containing the "fore-contrast" indicators, we put its previous subsentence s_{i-1} into d_{contrast}; for a subsentence containing the "post-contrast" indicators, we put the current subsentence s_i into d_{contrast}.

Finally, each document d in the training and test set is divided into three parts: d_{negation}, d_{contrast} and $d_{\text{no-shift}}$. Note that we only detected the explicit polarity shift in this step. In the next subsection, we shall employ a statistical method to detect the implicit sentiment inconsistency.

3.1.2. Statistical polarity shift detection for implicit contrasts

In this part, we propose a statistical method to detect the implicit sentiment inconsistency. The basic idea is based on the phenomenon that the sentiment inconsistency has the contrary polarity to that of its neighboring subsentences as well as

[1] In this work, we use the polarity shift trigger words collected in (Li et al., 2010) as a basic set, and add some hints based on our observations. Finally, the negators include "*no*", "*not*", "*n't*", "*none*", "*nobody*", "*nothing*", "*never*", "*hardly*", "*seldom*" and "*without*". The disjunctive conjunctions include "*but*", "*however*", "*yet*", "*unfortunately*", "*thought*", "*although*" and "*nevertheless*".

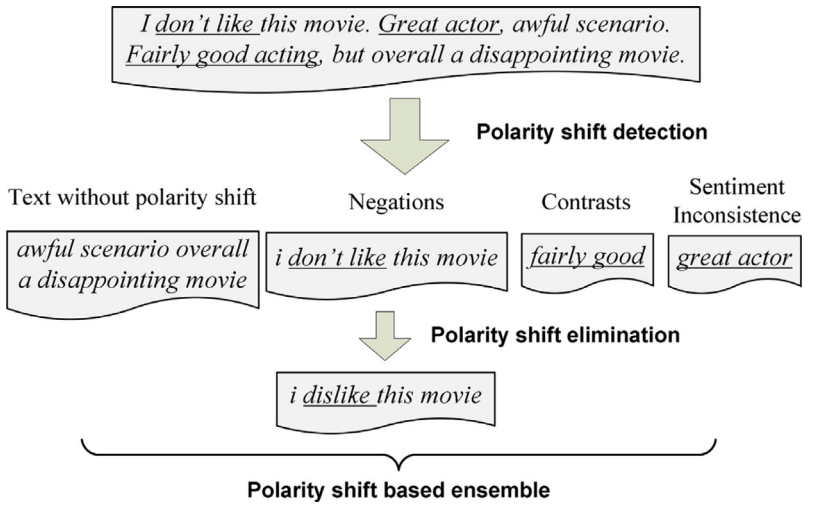

Fig. 1. An illustration of the Polarity Shift Detection, Elimination and Ensemble (PSDEE) approach.

1. **Input:** document $d = \{s_1, s_2, ..., s_m\}$, negation indicator set $\mathcal{N} = \{n_1, n_2, ...n_s\}$, and contrast indicator set $\mathcal{C} = \{c_1, c_2, ..., c_t\}$
2. **Output:** d_{negation}, d_{contrast}, $d_{\text{inconsistence}}$ and $d_{\text{no-shift}}$
3. for $i = 1, ..., m$:
4. for $j = 1, ..., |s_i|$:
5. if $w_{ij} \in \mathcal{N}$: put s_i into d_{negation}; # capture negations
6. continue;
7. if $w_{ij} \in \mathcal{C}_1$: put s_{i-1} into d_{contrast}; # capture contrasts (fore-contrast)
8. continue;
9. if $w_{ij} \in \mathcal{C}_2$: put s_i into d_{contrast}; # capture contrasts (post-contrast)
10. continue;
11. compute $r(w_{ij})$;
11. compute $h_(s_i)$;
12. if $h(s_i) < 0$: put s_i into $d_{\text{inconsistence}}$; #capture sentiment inconsistency
13. let $d_{\text{no-shift}} = d - d_{\text{negation}} - d_{\text{contrast}} - d_{\text{inconsistence}}$

Fig. 2. The pseudo-code of the hybrid polarity shift detection algorithm.

the whole review. Therefore, we could use a statistical method to detect the sentiment polarity of each subsentence, and compare the subsentence-level sentiment polarities with the sentiment polarity of the whole review. The subsentences that have the contrary sentiment polarities to the whole review are labeled as sentiment inconsistency. Line 11–12 in Fig. 2 shows the pseudo-code of the statistical method for detecting sentiment inconsistency.

Specifically, we employ the weighted log-likelihood ratio (WLLR) algorithm to detect inconsistent sentiment in the text. WLLR was designed as a feature selection method for text classification in (Nigam et al., 2000). Formally, WLLR measures the relevance of the feature t_i to the class label c_j as follows:

$$r(t_i,c_j) = p(t_i|c_j)\log\frac{p(t_i|c_j)}{p(t_i|\bar{c}_j)} \tag{1}$$

In our approach, we first use WLLR to obtain the relevance of each feature and two classes, i.e., Positive (+) and Negative (−), and define a WLLR score with regard to a feature t_i as:

$$r(t_i) = r(t_i,+) - r(t_i,-) \tag{2}$$

Second, we compute the orientation of each subsentence based on the WLLR score. Suppose a document d is composed of m subsentences $d = (s_1,s_2,\ldots,s_m)$, where each sentence s_i is represented by a list of words contained in the subsentence $s_i = (w_{i1},w_{i2},\ldots,w_{i|s_i|})$. We define the positive relevance and negative relevance of a subsentence s_i as

$$f(s_i) = \sum_{h-1}^{|s_i|} r(w_{ih}) \tag{3}$$

Finally, we define the sentiment inconsistency indicator functions as:

$$h(s_i) = y f(s_i) \tag{4}$$

where y is the class label. Note that for training document d_k, the class label y_k is already known (denoted by $y \in \{+1, -1\}$). For each test document $\tilde{d}_i$, we use the sum of relevance scores in the document as the approximate of the true class label:

$$\tilde{y} = sign\left(\sum_{j=1}^{m}\sum_{k=1}^{|s_j|} r(w_{jk})\right) = sign\left(\sum_{k=1}^{|\tilde{d}_i|} r(w_k)\right) \tag{5}$$

If $h(s_i) < 0$, (i.e., the sentiment polarity of s_i and d are different), we say, sentence s_i is sentiment inconsistent with the document d. Otherwise, we believe s_i does not have sentiment inconsistency.

3.2. Negation polarity shift elimination

In the second stage, we propose a polarity shift elimination algorithm to remove negations in the reviews. The idea is to use the antonym words to replace the negated words, such that the text in Review 1 "*I don't like this book*" is changed to "*I dislike the book*." An antonym dictionary is required in this process. In this part, we introduce a totally corpus-based method to construct a "corpus-sense" antonym dictionary, without using any lexical resources.

We use the WLLR metrics again to identify the most positive and negative features in the training corpus, choose adjectives, adverbs and verbs as candidate words, and rank the candidate words according to a decreasing order of $r(t_i,+)$ and $r(t_i,-)$ in Eq. (1), respectively:

$$\mathcal{W}_+ = [pw_1, pw_2, \ldots, pw_D], \tag{6}$$
$$\mathcal{W}_- = [nw_1, nw_2, \ldots, nw_D] \tag{7}$$

The antonym dictionary is then constructed by zipping $\mathcal{W}_+$ and $\mathcal{W}_-$. Each word pair in $\{(pw_i, nw_i)\}_{i=1}^{D}$ is considered as a pair of antonyms. It is important to notice that it is a corpus-sense antonym dictionary, rather than a common-sense antonym dictionary. For example we may learn an antonym word pair (*interesting*, *unrealistic*) from the Book review dataset, hence, the negation text "*the book is not interesting*" will be converted into "*the book is unrealistic*". It should be noted that although "*unrealistic*" is not a good opposite word of "*interesting*", the WLLR method can guarantee that "*unrealistic*" and "interesting" has the same level of sentiment strength based on learning from the corpus, and therefore will still make sense in sentiment classification.

3.3. The polarity shift based ensemble model

So far we have presented the (polarity shift detection approach and the negation polarity shift elimination approach. Each document in the training and test sets is split into three component parts: (1) the eliminated negation part, (2) the contrast part, and (3) the sentiment inconsistency part and (4) the part without polarity shift. In this subsection, we propose an ensemble model to train three component classifiers for sentiment classification, based on the abovementioned three parts of text, respectively.

The ensemble technique, that combines the outputs of several base classification models to form an integrated output, has become an effective classification method for many domains including sentiment classification (Xia & Zong, 2011; Xia, Zong, Hu, & Cambria, 2013; Xia, Zong, & Li, 2011). The pursuit of ensemble in this work is motivated by the intuition that an appropriate integration of different components with respect to polarity shift might leverage distinct strengths in sentiment classification. Let $g_{kj}(\mathbf{x})$ denote the output of the k-th base-classifier for the j-th class. Then, the weighted ensemble could be written by

$$f_j = \sum_{k=1}^{D} \theta_k g_{kj}(\mathbf{x}),\ j = +,- \tag{8}$$

where C and D are the number of classes and base-classifiers respectively, θ_k denotes the weights for each components. An effective ensemble system is supposed to assign a relatively larger weight to the base classifier trained on the polarity upshifted parts, while assign a relative smaller weight to the base classifiers trained on the polarity shifted parts.

In the stage of training, we therefore have four training-set components. On each component, we train a base sentiment classifier. In the stage of prediction, we in the same way separate a test review into four parts, each of which is predicted based on the corresponding training model. An ensemble of four component predictions is used as the final prediction. The ensemble weights of three component classifiers are learnt by the stacking algorithm (Džeroski & Ženko, 2004) In the stacking framework, the probabilistic outputs of all base classifiers are used as meta-learning features, and a leave-one-out procedure is applied to the training data to train the ensemble weights in meta-learning.

4. Experiments

4.1. Datasets and settings

We systematically evaluate our approach on the Multi-domain sentiment datasets, which were introduced by Blitzer, Dredze, and Pereira (2007) and have been widely used in sentiment classification. It consists of four domains (i.e., Book, DVD, Electronics, and Kitchen) of reviews extracted from Amazon.com. Each of the four datasets contains 1000 positive reviews and 1000 negative reviews.

We evaluate three classification algorithms, i.e., linear SVM, logistic regression and Naïve Bayes. For linear SVM, we use the LibSVM[2] toolkit. For logistic regression, we use the LibLinear[3] toolkit. For Naïve Bayes, we use the OpenPR-NB toolkit (Xia et al., 2011). Following the mostly-used settings in sentiment classification, we examine two kinds of features, i.e., unigrams, and both unigrams and bigrams. The standard bool value (i.e., presence of features) is used as the feature weighting scheme because it was reported in (Pang et al., 2002) that Bool term weighting performed better than TF and TF-IDF in sentiment classification. All of the experiments are conducted by a 5-fold cross validation procedure. The following results are reported in terms of an average accuracy. The paired *t*-test (Yang & Liu, 1999) is performed for significant testing with a default significant level at 0.05.

4.2. Compared systems

We implement four related methods that tackle the polarity shift problem for document-level sentiment classification, and compare our PSDEE approach with them.

(1) **Baseline** the standard machine learning methods based on BOW representation without handling polarity shift.
(2) **DAS** the method proposed by Das (2001), where "*NOT*" is added to the sentiment words in the scope of negation, e.g., "*The book is not interesting*" is converted to "*The book is interesting-NOT*".
(3) **REV** a method similar to System 1, but the sentiment words in the scope of negation are reversed to their antonyms, e.g., "*The book is not interesting*" is converted to "*The book is boring*".
(4) **LSS** the method proposed by Li et al. (2010), where each text is separated into a polarity-shifted part and a polarity-unshifted part, based on which two component classifiers are trained and combined for sentiment classification.
(5) **PSDEE** our approach based on polarity shift detection, elimination and ensemble.

4.3. Comparison with alternative methods

From Tables 1–3, we report the classification accuracy of five systems using unigrams, and both unigrams and bigrams, based on three classification algorithms. We observe two types of features respectively.

4.3.1. Unigram features

As shown in Table 1, it is easy for us to observe that DAS approach yields very limited improvements (1% in average) compared to the baseline system. This result confirms the reports in (Pang & Lee, 2008). It suggests that simply appending "*NOT*" to the words in the negation scope is not effective in addressing the polarity shift problem. The REV approach also yields very limited performance. Although polarity shifts in negations are removed, many other types of polarity shifts are not considered. Therefore, the increase of classification accuracy is slight. The LSS approach outperforms the baseline system by 1.5, 3.1, 1.6 and 2.2 percentages on four datasets respectively. The improvements are very limited. Compared to previous three systems, our PSDEE approach gains much more improvements. It increases the baseline performance by 5.9, 4.6, 3.4 and 3.5 percentages, and outperforms Li's approach by 4.4, 1.5, 1.8 and 1.3 percentages, on four datasets respectively. Observing the results of the other two classifiers in Tables 2 and 3, we could find that the DAS and REV approaches still achieve slight improvements (less than 1%). The improvements of LSS are still limited. It improves the baseline performance by 2.0% and 1.2% on logistic regression and Naïve Bayes, respectively. By contrast, our PSDEE approach outperforms the baseline system by 3.4% and 1.6%, and outperforms the LSS approach by 1.4% and 0.4% on average on the two classifiers respectively. All improvements are significant according to the paired *t*-test, except for the results of Naïve Bayes compared to LSS approach.

4.3.2. Both unigram and bigram features

By observing Table 1 again, we can find that compared to using unigrams, the accuracy of each system using both unigrams and bigrams is also improved. It is also worth noting that because using bigram features already captures a part of negations (i.e., "*don't like*"), it is more difficult to gain improvements of addressing polarity shift in this case. Nevertheless, comparing different systems, we could draw similar conclusions as that in unigram features. In comparison with the baseline

[2] http://www.csie.ntu.edu.tw/~cjlin/libsvm/.
[3] http://www.csie.ntu.edu.tw/~cjlin/liblinear/.

Table 1
Performance of different systems using linear SVM.

Datasets	Unigram features					Unigram and bigram features				
	Baseline	DAS	REV	LSS	PSDEE	Baseline	DAS	REV	LSS	PSDEE
Book	0.745	0.763	0.754	0.760	**0.804**	0.775	0.777	0.785	0.788	**0.814**
DVD	0.764	0.771	0.774	0.795	**0.810**	0.791	0.793	0.797	0.809	**0.820**
Electronics	0.796	0.813	0.804	0.812	**0.830**	0.818	0.834	0.83	0.841	**0.848**
Kitchen	0.822	0.820	0.831	0.844	**0.857**	0.847	0.844	0.854	0.870	**0.871**
Avg.	0.782	0.792	0.791	0.803	**0.825**	0.808	0.812	0.817	0.827	**0.838**

Table 2
Performance of different systems using logistic regression.

Datasets	Unigram features					Unigram and bigram features				
	Baseline	DAS	REV	LSS	PSDEE	Baseline	DAS	REV	LSS	PSDEE
Book	0.771	0.775	0.774	0.784	**0.809**	0.779	0.789	0.790	0.809	**0.823**
DVD	0.785	0.800	0.790	0.815	**0.819**	0.801	0.802	0.807	0.823	**0.832**
Electronics	0.803	0.815	0.817	0.823	**0.834**	0.827	0.833	0.837	0.844	**0.853**
Kitchen	0.834	0.841	0.836	0.851	**0.866**	0.851	0.858	0.861	0.872	**0.874**
Avg.	0.798	0.808	0.804	0.818	**0.832**	0.815	0.821	0.824	0.837	**0.846**

Table 3
Performance of different systems using Naïve Bayes.

Datasets	Unigram features					Unigram and bigram features				
	Baseline	DAS	REV	LSS	PSDEE	Baseline	DAS	REV	LSS	PSDEE
Book	0.779	0.7830	0.784	0.792	**0.796**	0.811	0.815	0.817	0.822	**0.829**
DVD	0.795	0.7930	0.798	0.810	**0.813**	0.824	0.826	0.8332	0.837	**0.837**
Electronics	0.815	0.8275	0.826	0.824	**0.825**	0.841	0.857	0.8545	0.852	**0.866**
Kitchen	0.830	0.8475	0.838	0.840	**0.850**	0.878	0.879	0.877	0.883	**0.891**
Avg.	0.805	0.813	0.812	0.817	**0.821**	0.839	0.844	0.845	0.849	**0.856**

system, the improvements of DAS and REV are still very limited (0.4% and 0.9% on average). According to the average results, the LSS approach improves the baseline performance by 1.9%, and our PSDEE approach outperforms the Baseline and LSS systems by 3.0% and 1.1% respectively. In Tables 2 and 3, we could observe that DAS and REV still have negligible effects. The LSS approach outperforms the baseline system by 2.2% and 1.0% on the two classifiers respectively. On the contrary, in logistic regression, our PSDEE approach outperforms the baseline system and LSS approach by 3.1% and 0.9% respectively; in Naïve Bayes, our PSDEE system improves the average score by 1.7% and 0.7% compared with the baseline and LSS systems. All improvements are significant according to the paired *t*-test.

5. Discussions

In this section, we will provide in-depth discussion by investigating the performance of different settings at each stage of our PSDEE approach. For simplicity, we only present the experimental results on linear SVM. Similar conclusions could be drawn in case of logistic regression and Naïve Bayes.

5.1. The effect of hybrid polarity shift detection

In order to evaluate the effect of the polarity shift detection method used in our model, in this part, we compare three polarity shift detection methods proposed in Section 3.1.

(1) **Only Rule** where polarity shifts are detected only based on rules.
(2) **Only Statistic** where polarity shifts are detected only based on statistical method.
(3) **Hybrid Detection** which employs rules and statistical method together for detecting polarity shift.

In Table 4, we report the sentiment classification performance of PSDEE by using three different polarity detection methods, as well as the baseline performance.

As for unigrams, we can observe that compared to the baseline system, the improvements of **Only Rule** and **Only Statistic** systems are 2.7% and 3.4%, and **Hybrid Detection** system outperforms the baseline system by 4.3% on average. As for both

unigrams and bigrams, all the three systems yield better performances than the baseline system. Furthermore, comparing the experimental results of the three systems, we can find that although **Only Statistic** outperforms **Only Rule** by 0.7% and 0.8% on the two kinds of features, and **Only Rule** performs better than **Only Statistic** on DVD and Electronics with unigram features. Again, we find that the performance of the **Hybrid Detection** system performs better than the other two systems across different settings. In general, we draw the conclusion that there is no consistent winner between **Only Rule** and **Only Statistic**, while **Hybrid Detection** system is robustly the best.

5.2. The effect of negation polarity shift elimination

In this subsection, we focus on evaluating the effect of the second stage of our PSDEE model (i.e., negation polarity shift elimination (PSE)). In Table 5, we compare the classification accuracy before and after PSE on two classification schemes: **Single Classifier** and **Ensemble Classifier**. Specifically, we use the polarity shift detection to separate the training and test set into four subsets. Here, single classifier denotes one single classifier trained on a joint set of four subsets. The difference from Single and Baseline is that the negations have been eliminated in the Single model. Ensemble Classifier denotes a weighted combination of each component classifiers.

As can be seen in Table 5, for using unigram features, the results after PSE outperforms that before PSE by 0.9% and 1.2% on Single and Ensemble Classifiers respectively. As for both unigrams and bigrams, in comparison with the results before PSE, the improvements after PSE are 0.9% on both classifiers. The results are reasonable since with the elimination of negation polarity shift, the new sentiment words, which replace the original sentiment words in the negation scope, will correct some learning and prediction errors caused by negations.

5.3. The effect of polarity shift based ensemble

In Table 6, we report the classification performance of each base classifier, as well as the performance of a single classifier and ensemble classifier. As we can see from Table 6, the ensemble model exceeds the single classifier significantly according to the paired *t*-test.

In Table 7, we furthermore report the average weights of each component trained in stacking. We can find that the weight of the No-shift component is the largest in four parts, the weight for the eliminated negation component is second largest, and the weights for the explicit contrast and sentiment inconsistency part are relatively smaller. This confirms our motivation in Section 3.2 that an appropriate ensemble by assigning larger weight to polarity unshifted text while assigning smaller weight to the polarity shifted text, will enhance the sentiment classification performance.

6. Related work

Sentiment analysis is a recent and explosively growing research field, widely employed by the industry for diverse applications such as data visualization (Cambria et al., 2010), human-computer interaction (Poria, Cambria, Hussain, & Huang, 2015), and e-health (Cambria, Hussain, Havasi, & Eckl, 2010b). The task of automatically extracting polarity from text, however, is very difficult as it entails known NLP problems such as anaphora resolution, sarcasm detection, and polarity shift.

In this work, we focus on reviewing methods for addressing the polarity shift problem at the document-level sentiment classification. In the literature, methods of document-level sentiment classification can be separated as term-counting methods and machine learning methods. In term-counting methods, the overall sentiment of a text is obtained by summing up the scores of content words or phrases in the text, according to manually-collected or external lexical dictionaries (Turney, 2002). The machine learning methods represents a piece of text by a BOW model and then use standard machine learning algorithms for classification (Pang et al., 2002).

The way to handle polarity shift also differs in two types of methods accordingly. It is relatively easy to encode polarity shift in the term-counting methods For example, we can reverse the sentiment of the polarity-shifted words and phrases directly, and then sum up the sentiment score word by word (Hu & Liu, 2004; Kennedy & Inkpen, 2006; Polanyi & Zaenen, 2004). Wilson et al. (2005) discussed some complex negation effects by using conjunctive and dependency relations among polarity words. Li et al. (2013) developed four rules (intra-clause rule, intra-sentence rule, extra-sentence rule, and extra-paragraph rule) for detecting different types of polarity shift, and employed a term-counting model to encode the information of polarity shifts. Empirical studies showed that their method yield much better performances than the basic term-counting approach.

Kennedy and Inkpen (2006) handled polarity shift in both term-counting and machine learning applications. However, although the system is effective in term-counting systems, it is difficult to outperform the baselines in machine learning methods. It is relatively difficult to handle polarity shift in machine learning methods, perhaps because the polarity shift information is hard to be integrated into the BOW model. Das and Chen (2001) proposed a simple method by adding "NOT" to sentiment words. But Pang et al. (2002) reported that the effect is very limited in improving the sentiment classification accuracy. Some researchers attempted to model polarity shift by conducting more complex linguistic analysis. For example, Na et al. (2004) tried to model negation by investigating specific part-of-speech tag patterns. Kennedy and Inkpen (2006) employed syntactic parsing to capture three types of valence shifters (negative, intensifiers, and diminishers).

Table 4
Comparison of three polarity shift detection methods.

Datasets	Unigram features				Unigram and bigram features			
	Baseline	Only rule	Only statistic	Hybrid detection	Baseline	Only rule	Only statistic	Hybrid detection
Book	0.745	0.781	0.801	**0.804**	0.775	0.796	0.812	**0.814**
DVD	0.764	0.797	0.790	**0.810**	0.791	0.810	0.814	**0.820**
Electronics	0.796	0.824	0.820	**0.830**	0.818	0.836	0.841	**0.848**
Kitchen	0.822	0.836	0.853	**0.857**	0.847	0.857	0.868	**0.871**
Avg.	0.782	0.809	0.816	**0.825**	0.808	0.825	0.833	**0.838**

Table 5
The comparison of systems performance before and after polarity shift elimination.

Datasets	Unigram features				Unigram and bigram features			
	Single classifier		Ensemble classifier		Single classifier		Ensemble classifier	
	Before PSE	After PSE	Before PSE	After PSE	Before PSE	After PSE	Before PSE	After PSE
Book	0.746	0.754	0.791	**0.804**	0.775	0.785	0.798	**0.814**
DVD	0.764	0.774	0.795	**0.810**	0.791	0.797	0.819	**0.820**
Electronics	0.796	0.804	0.822	**0.830**	0.818	0.830	0.837	**0.848**
Kitchen	0.822	0.831	0.846	**0.857**	0.847	0.854	0.862	**0.871**
Avg.	0.782	0.791	0.813	**0.825**	0.808	0.817	0.829	**0.838**

Table 6
The performance of single classifier and Ensemble Classifier.

Dataset	Unigram features			Unigram and bigram features		
	Baseline	Single	Ensemble	Baseline	Single	Ensemble
Book	0.7455	0.754	0.7915	0.775	0.785	0.798
Dvd	0.764	0.774	0.7955	0.7905	0.797	0.8195
Electronics	0.796	0.804	0.822	0.8185	0.830	0.837
Kitchen	0.822	0.831	0.846	0.847	0.854	0.862
Avg.	0.782	0.791	0.814	0.808	0.817	0.829

Table 7
The component weights of different parts with respect to polarity shift.

	No-shift	Negation (Eliminated)	Explicit contrast	Sentiment inconsistency
Unigram features	0.40	0.22	0.17	0.21
Unigram and bigram features	0.55	0.21	0.10	0.14

Ikeda et al. (2008) proposed a method in the machine learning framework to model polarity-shifters for both word-wise and sentence-wise sentiment classification, based on an extra dictionary extracted from General Inquirer. Li and Huang (2009) first classify each sentence in a text into a polarity-unshifted part and a polarity-shifted part according to certain rules, and then represent them as two bags of words for sentiment classification. Li et al. (2010) further proposed a method to separate the shifted and unshifted text based on training a binary detector. An ensemble of two component classifiers is used at last to get the final sentiment polarity. Orimaye, Siew (2012) proposed a Sentence Polarity Shift (SPS) algorithm that employs three polarity shift patterns to extract sentences with consistency sentiments in a review and removes the inconsistent ones for sentiment classification.

This work resembles the work by Li et al. (2010) in the manner of polarity shift separation and ensemble. We update their method by two aspects. First, we employ a hybrid polarity shift detection algorithm for different types of polarity shifts. Second, we add propose a new stage in the model to eliminate polarity shift in negations.

7. Conclusion

The work describes a cascade model, namely Polarity Shift Detection, Elimination and Ensemble (PSDEE), to address the polarity shift problem in document-level sentiment analysis. In the first stage, we propose a hybrid model that employs both rule-based and statistic-based methods to detect different types of polarity shifts. Specifically, we use a rule-based method to detect explicit negations and contrasts, and a statistical method to detect the implicit sentiment inconsistency. In the second stage, we introduce a novel method called antonym reversion to eliminate polarity shifts in negations. After the first two stages, a piece of text is separated into four subsets, namely the polarity-unshifted text, eliminated negations, explicit contrasts and sentiment inconsistency. In the third stage, a weighted ensemble of base classifiers trained on component text

subsets is employed as the final sentiment classifier, with the aim to leverage text with different types of polarity shifts. We conduct a range of experiments including four sentiment datasets, three classification algorithms and two types of features. The results demonstrate the effect of our PSDEE approach compared to several relate work that addresses polarity shift in document-level sentiment classification.

Acknowledgement

The work is supported by the Natural Science Foundation of China (61305090 and 61272419), the Jiangsu Provincial Natural Science Foundation of China (BK2012396), and the Research Fund for the Doctoral Program of Higher Education of China (20123219120025).

References

Blitzer, J., Dredze, M., & F. Pereira. (2007). Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th annual meeting of the association of computational linguistics* (pp. 440–447).

Cambria, E., Hussain, A., Havasi, C., & Eckl, C. (2010). Sentic computing: Exploitation of common sense for the development of emotion-sensitive systems. *LNCS, 5967*, 148–156 Springer.

Cambria, E., Hussain, A., Havasi, C., & Eckl, C. (2010). SenticSpace: Visualizing opinions and sentiments in a multi-dimensional vector space. *LNAI, 6279*, 385–393 Springer.

Cambria, E. Hussain, A. Durrani, T. Havasi, C. Eckl, C. Munro J. (2010). Sentic computing for patient centered applications. In *IEEE ICSP* (pp. 1279–1282). Beijing.

Cambria, E., Mazzocco, T., Hussain, A., & Eckl, C. (2011). Sentic medoids: Organizing affective common sense knowledge in a multi-dimensional vector space. *LNCS, 6677*, 601–610.

Cambria, E., Schuller, B., Liu, B., Wang, H., & Havasi, C. (2013). Knowledge-based approaches to concept-level sentiment analysis. *IEEE Intelligent Systems, 28*(2), 12–14.

Cambria, E., Schuller, B., Liu, B., Wang, H., & Havasi, C. (2013). Statistical approaches to concept-level sentiment analysis. *IEEE Intelligent Systems, 28*(3), 6–9.

Choi, Y., & Cardie, C. (2008). Learning with compositional semantics as structural inference for subsentential sentiment analysis. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)* (pp. 793–801).

Das, S., & Chen, M. (2001). Yahoo! for Amazon: Extracting market sentiment from stock message boards. In *Proceedings of the Asia pacific finance association annual conference*.

Džeroski, S., & Ženko, B. (2004). Is combining classifiers with stacking better than selecting the best one? *Machine Learning, 54*, 255–273.

Gangemi, A., Presutti, V., & Reforgiato, D. (2014). Frame-based detection of opinion holders and topics: A model and a tool. *IEEE Computational Intelligence Magazine, 9*(1), 20–30.

Hu, M. & Liu, B. (2004). Mining and Summarizing Customer Reviews. In KDD.

Ikeda, D., Takamura, H., Ratinov, L., Okumura, M. (2008). Learning to shift the polarity of words for sentiment classification. In *Proceedings of the international joint conference on natural language processing (IJCNLP)*.

Kennedy, A., & Inkpen, D. (2006). Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence, 22*, 110–125.

Li, S., & Huang, C. (2009). Sentiment classification considering negation and contrast transition. In *Proceedings of the pacific Asia conference on language, information and computation (PACLIC)*.

Li, S., Xia, R., Zong, C., & Huang, C. (2009). A framework of feature selection methods for text categorization. In *Proceedings of the annual meeting of the association for computational linguistics and international joint conference on natural language processing (ACL/IJCNLP)* (pp. 692–700).

Li, S., Lee, S., Chen, Y., Huang, C., & Zhou, G. (2010). Sentiment classification and polarity shifting. In *Proceedings of the international conference on computational linguistics (COLING)*.

Li, S., Wang, Z., Li, S., & Huang, C. (2013). Sentiment classification with polarity shifting detection. In *Proceedings of the international conference on Asian language processing*.

Liu, B. (2012). Sentiment analysis and opinion mining. *Morgan & Claypool*.

Na, J., Sui, H., Khoo, C., Chan, S., & Zhou, Y. (2004). Effectiveness of simple linguistic processing in automatic sentiment classification of product reviews. In *Proceedings of the conference of the international society for knowledge organization*.

Nigam, K., McCallum, A., Thrun, S., & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning, 39*(2/3), 103–134.

Orimaye, S., Alhashmi, S., & Siew, E. (2012). Buy it – Do not buy it: sentiment classification on Amazon reviews using sentence polarity shift. In *Proceedings of the pacific rim international conference on artificial intelligence (PRICAI)*.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)* (pp. 79–86).

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 2*(1–2), 1–135.

Polanyi, L., Zaenen, A. (2004). Contextual lexical valence shifters. In *Proceedings of the AAAI spring symposium on exploring attitude and affect in text*.

Poria, S., Cambria, E., Hussain, A., & Huang, G.-B. (2015). Towards an intelligent framework for multimodal affective data analysis. *Neural Networks, 63*, 104–116.

Poria, S., Gelbukh, A., Cambria, E., Hussain, A., & Huang, G.-B. (2014). EmoSenticSpace: A novel framework for affective common-sense reasoning. *Knowledge-Based Systems, 69*, 108–123.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics, 37*(2), 267–307.

Turney, P. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the annual meeting of the association for computational linguistics (ACL)*.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*.

Xia, R., & Zong, C. (2011). A POS-based ensemble model for cross-domain sentiment classification. In *Proceedings of the international joint conference on natural language processing (IJCNLP)*.

Xia, R., Xu, F., Zong, C., Li, Q., Qi, Y., & Li, T. (2015). Dual sentiment analysis: Considering two sides of one review. *IEEE Transactions on Knowledge and Data Engineering, 27*(8), 2120–2133.

Xia, R., Zong, C., Hu, X., & Cambria, E. (2013). Feature ensemble plus sample selection: Domain adaptation for sentiment classification. *IEEE Intelligent Systems, 28*(3), 10–18.

Xia, R., Zong, C., & Li, S. (2011). Ensemble of feature sets and classification algorithms for sentiment classification. *Information Sciences, 181*(6), 1138–1152.

Yang, Y., & Liu, X. (1999). A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval* (pp. 42–49).