

Sentiment Classification using Subjective and Objective Views

Suke Li

SSM, Peking University
Yiheyuan Road 5, Beijing,
China, 100871

ABSTRACT

This work proposes a new semi-supervised sentiment classification method by exploiting a large number of unlabeled instances to conduct sentiment classification for Web consumer reviews. In the proposed method every consumer review has two views: subjective view and objective view. The subjective view of a consumer review reflects the opinions expressed by opinion words, while the objective view is constructed by the remaining text features. This work is trying to combine two kinds of views to carry out sentiment classification. The method is based on the co-training framework which needs three basic sentiment classifiers to iteratively get the final sentiment classifier. In the proposed method, the first sentiment classifier is constructed using the common unigram features coming from consumer reviews. The second sentiment classifier is trained on the subjective views constructed by opinion words extracted from consumer reviews. The remaining text features of these reviews are used for obtaining the objective views which can be trained for the third classifier. Experimental results show the proposed method is effective, and it has better performance than the Self-learning SVM method.

General Terms

Algorithm

Keywords

Sentiment Classification, Opinion Mining, Text Mining, Web Mining, Data Mining.

1. INTRODUCTION

Within recent years, Web-based applications now deliver complex and varied functionality to Web users. The interactions between common Web users and Web systems become very easy when these Web systems are openly accessed in a social network ecosystem. Especially, user-generated content including Web reviews is widely spread in Web forums, social networks, and some e-commerce Web sites. Web users are used to publishing product reviews after they have purchased products or services. At the same time, Web users now increasingly depend on the published reviews by other people before they make purchasing decisions. Web reviews significantly affect the activities of consumers because of their potential economic values. Therefore mining knowledge from Web reviews becomes an important part of many intelligence systems such as recommendation systems. Sentiment classification is one main task of opinion mining. Plenty of research publications have focused on sentiment classification. The approaches of sentiment classification can roughly fall into two basic categories. The methods in the first category rely on language resources which are constructed before carrying out sentiment classification. The language resources include sentiment lexicons and natural language corpus libraries. Researchers usually use some natural language processing

techniques combined with language resources to improve the accuracy of the sentiment classification. The methods in the second category try to employ machine learning to do sentiment classification. Machine learning-based sentiment classification methods contain supervised and semi-supervised methods which need some training instances to learn to get the final sentiment classifiers. If we just simply apply sentiment classifier trained by instances coming from one domain on the other domain, it could be a hard task to get high classification accuracy because different domains might have different sentiment features or opinion words. It could be a hard task to label training instances for each domain. However, a semi-supervised-based method only requires a small number of training examples, and it exploits useful information extracted from a large number of unlabeled instances to boost the sentiment classification.

Most Helpful Customer Reviews

721 of 747 people found the following review helpful

★★★★★ Update - Now has Amazon VOD Support June 16, 2012

By G. Peck

Size Name: 32-Inches | Amazon Verified Purchase

As many others have pointed out, this set has great picture quality. However, you will really have to work at it. Out of the box the picture looked awful and I was ready to return this set. Then I was lucky enough to find a review by "Gordon" for the Samsung UN32EH5000. I used the exact settings he specified and now the picture looks great.

Fig 1: A product review example

The basic observation shows a product review may include both subjective and objective information. Subjective information indicates the opinions of opinion holders, while objective texts show some objective facts. Web users are likely to express their opinions using some opinion words. For example two opinion words "great" and "delicious" in the sentence "The food is great and delicious." These opinion words are subjective. Subjective texts can have positive or negative sentiment. Positive sentiment expresses happy, supportable, and positive sentiment, and negative sentiment reflects angry, opposite, and depressive sentiment. Figure 1 is an example of product review coming from Amazon¹. The review has some subjective opinion expressions such as "awful" and "great". These opinion words reflect the subjective feeling of the opinion holder. However, not every word is subjective. Most words of the review give objective information.

To simplify the problem, extracted opinion words from reviews are looked as subject information, the remaining text of reviews is looked as objective information. A consumer review has two kinds of views: one is subjective view, and the other is object view. Opinion words with sentiment polarities have contribution to construct subjective views. Because opinion words in the subjective view can be used to express sentiment information, so opinion words are looked as subjective features, while other

¹ <http://www.amazon.com>

words in the objective view are classified into objective features.

This work proposes a semi-supervised sentiment classification method based the co-training algorithm [1]. The method only requires a small number of training instances and a large number of unlabeled instances. The proposed method is based co-training algorithm framework which needs three sentiment classifiers. The first sentiment classifier uses all subjective and objective features to train a sentiment classifier. The second sentiment classifier is constructed by opinion words from subjective views. The third sentiment classifier derives from objective features. The proposed method firstly gets three sentiment classifiers on training data set, and then the three sentiment classifiers iteratively co-train together on the unlabeled data set to find the most possible rightly classified instances and put these instances into the training data set. The iterative process continues until no instance is left in the unlabeled data set. Experimental results show the proposed method outperforms the Self-learning SVM method.

The rest of the paper is organized as follows. In section 2, this work shows the related works about sentiment classification. Section 3 gives the semi-supervised sentiment classification approach. Experimental results and conclusions are presented in section 4 and section 5 respectively.

2. PRELATED WORKS

Opinion mining has become an attract research area in the field of data mining. Sentiment classification is the main task of opinion mining. Sentiment classification has drawn much attention in recent years. The early publication addressing the sentiment classification is Hatzivassiloglou and McKeown's work [2] which gives the method of predicting sentiment orientations of conjoined adjectives. Turney [3] proposed PMI (Pointwise Mutual Information) method to determine the sentiment orientation of words. Turney [4] also combined PMI with information retrieval to conduct document sentiment classification. Pang [5] firstly employed machine learning techniques such as Naive Bayes, maximum entropy and support vector machines to carry out sentiment classification on movie reviews, and experimental results proved that SVM-based method was effective. Pang [6] also gave an important survey about sentiment classification. Although SVM is successful in sentiment classification, Moraes et al. [7] showed Artificial Neural Networks (ANN) outperformed SVM to some extend. Ensemble technique by integrating different feature sets and classification algorithms is also proved to be useful for sentiment classification. Xia et al. [8] made a comparative study of the effectiveness of ensemble technique for sentiment classification. Besides various supervised machine learning methods in the same domain, there are some publications that focus on cross-domain sentiment classification, such as [9] [10] [11] [12]. It is also useful to combine machine learning and lexicon-based methods together to do sentiment classification, such as Qiu et al.'s work [13].

The proposed method is based two kinds of views: subject view and objective view. There are some research works that mention how to extract subjectivity. Sarvabhotla et al. [14] used statistical methodology combining with feature selection methods to extract subjectivity for sentiment classification.

Some publications focus on semi-supervised sentiment classification. Zhou [15] gave the semi-supervised machine learning sentiment classification method based on the active deep network. Li et al. [16] conducted sentiment classification based on co-training framework combining two views: one view was "personal view", the other view was "impersonal view".

The previous work [17] also gave a spectral clustering-based semi-supervised sentiment classification method. In the work [18], the authors proposed a linear regression with rules-based approach to ranking product features according to their importance.

3. THE PROPOSED SCHEME

Figure 2 gives some details about the proposed method. In this method, only a small number of labeled instances are used as the initial training instances. A large number of instances in the unlabeled data set are used to improve sentiment classification. The method is based co-training [1] algorithm which is a semi-supervised self-boost framework. A review has two views: the subject view and the objective view. Opinion words extracted from a review comprise of a subjective view, while the objective view is constructed by the remaining text features. All text features which are the frequencies of unigrams in the training data set are used to train SVM to get the first sentiment classifier. The proposed method relies on SVM (Support Vector Machine) to do sentiment classification. After the preprocessing action, TinySVM² is used o get the first sentiment classifier. Opinion words in subjective views are used to train the second sentiment classifier. Non-opinion words in objective views are training features of the second sentiment classifier. The three sentiment classifiers co-train together on the training data set to classify the instances in the unlabeled data set. The algorithm finds the most possible rightly classified instances in the unlabeled data set, and then put these instances into training data set again. The co-training process continues until there is no instance left in the unlabeled data set.

So far, opinion words from reviews must be extracted to construct subject views. Web users trend to use simple adjectives to describe products to express their opinions. While, the early research indicates adjectives can have rich sentiment information. For example, in the sentence "The food is delicious", the adjective "delicious" has positive sentiment. Sometime an adjective appears with a negative indicator, for instance, "The food is not bad", "not" is a negative indicator. Opinion word "bad" has negative sentiment, but "not bad" has positive sentiment. To simplify the problem, if an opinion word appears together with a negative indicator, the opinion word and the negative indicator together are looked as a single opinion word. Adjectives with POS (Part-of-Speech) labels of JJ, JJR, and JJS are extracted as opinion words.

These negative indicators, just as in the previous work [17] [18], including "not", "no", "donot", "do not", "didn't", "did not", "was not", "wasn't", "isn't", "isn't", "weren't", "weren't", "doesn't", "doesn't", "hardly", "never", "neither", and "nor".

² <http://chasen.org/~taku/software/TinySVM>

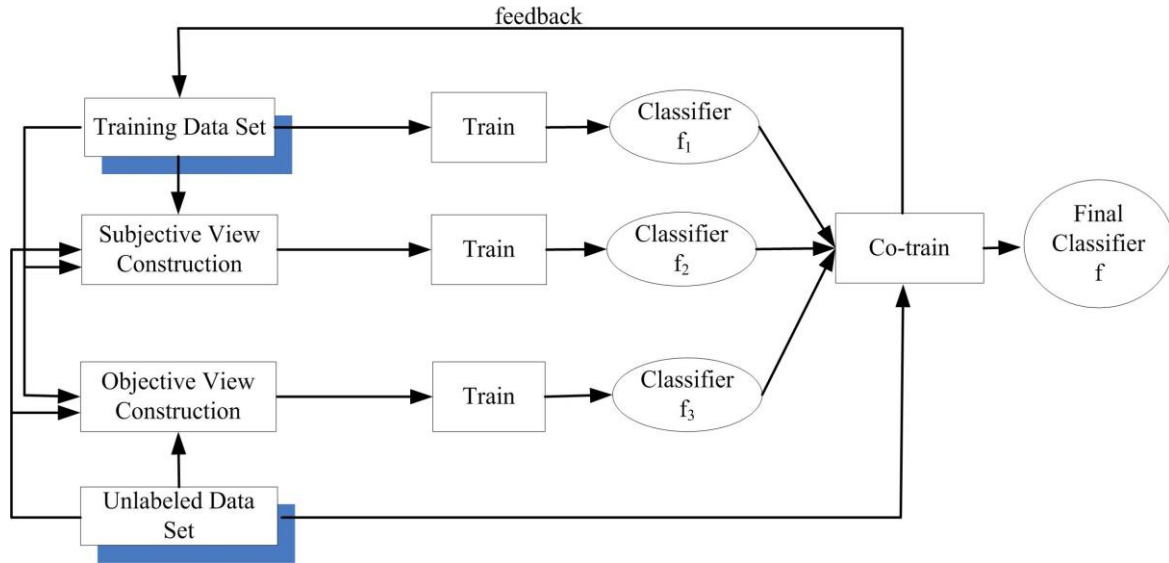


Fig 2: Overview of the proposed method

Algorithm 1 Two-View Co-training SVM Method

Input: Training data set $T = \{t_1, t_2, \dots, t_x\}$, T is balanced data set; Unlabeled data set $U = \{u_1, u_2, \dots, u_y\}$;

Output: Sentiment classifier C ;

- 1: while there are left unlabeled instances in the unlabeled data set do
- 2: Use training data set T and SVM to get the first sentiment classifier f_1 ;
- 3: Use opinion words (subjective views) extracted from the training instances to train SVM to get the second classifier f_2 ;
- 4: Use non-opinion text features (objective views) from training data set and SVM to get the third sentiment classifier f_3 ;
- 5: Use f_1 to classifier instances in U , and get positive instance set P_{f1} and Negative instance set N_{f1} (each iteration takes the most 50 possible classified instances);
- 6: Use f_2 to classify unlabeled data set U , and get positive instance set P_{f2} and negative instance set N_{f2} (each iteration takes the most 50 possible classified instances);
- 7: Use f_3 to classify unlabeled data set U , and get positive instance set P_{f3} and negative instance set N_{f3} (each iteration takes the most 50 possible classified instances);
- 8: $T = T \cup P_{f1} \cup N_{f1} \cup P_{f2} \cup N_{f2} \cup P_{f3} \cup N_{f3}$;
- 9: end while;
- 10: Get the final classifier $C = f_i$;
- 11: return C .

If there is a negative indicator in the context window $[-3, 0]$ of an opinion word, then the negative indicator combined with the sentiment word together are looked as a single opinion word. The context of a word window $[-3, 0]$ means the left distance coverage of the word in a clause is 3 words from the opinion

word. The detailed steps of the proposed method are shown as Algorithm 1 (Two-View Co-training SVM Method) which is the co-training-based framework [1]. The co-training framework is an effective semi-supervised learning framework which needs use different views in the data set to construct sentiment classifier. The Two-view Co-training SVM method has iteratively selected classified instances from the unlabeled data set into training space. In each iteration, every sentiment classifier selects the most 50 possible instances from the unlabeled data set, until the unlabeled data set is empty.

4. EXPERIMENTS

4.1 Experiment Data

The phone and laptop reviews are crawled from Amazon³ and the hotel reviews from Tripadvisor⁴ respectively. Consumer reviews were extracted from these Web pages. OpenNLP⁵ was employed to conduct some shallow nature language processing to get sentences and POS tags. Each data category includes 1000 positive reviews and 1000 negative reviews, as Table 1 shows.

Table 1. Experimental statistics

Data set	# of positive reviews	# of negative reviews	# of sentences
phone	1000	1000	28811
laptop	1000	1000	14814
hotel	1000	1000	18694

These data sets also were studied in the previous publications [17] [18]. Reviews were also segmented into sentences. For instance, the laptop reviews are segmented into 14814 sentences. The sentiment polarity of a review in a training data

³ <http://www.amazon.com>

⁴ <http://www.tripadvisor.com/>

⁵ <http://opennlp.apache.org/>

set is assigned according to their review ratings. A review rating is a real number ranging from 1 to 5. When a review rating is greater than 3, then the review is a positive review; when a review rating is less than 3, the review is a negative review.

4.2 Compared Method

The earliest use of SVM (Support Vector Machine) [19] to determine the sentiment polarity of consumer reviews is Pang's research work [5]. It is believed SVM is one of the best text classification approaches. Because the proposed method is a semi-supervised sentiment classification method, it is proper to compare the proposed method with the Self-learning SVM method. Therefore, Self-learning SVM method is the baseline. We use the frequencies of unigrams of reviews as the training and test features for SVM. TinySVM⁶ software is used to conduct SVM-based sentiment classification. The Self-learning SVM is a bootstrap approach to learning as Algorithm 2 shows. The method is also used in the previous work [17] to do some comparisons. The algorithm iteratively selects the most likely correctly classified instances which are determined by their distances to classification hyperplane, and put them into the training data set. A new sentiment classifier is built by training on the new training data set. Repeat the above steps until there are no unlabeled reviews left to be added into the training data set so far. When a classified review has greater distance from SVM hyperplane, the review is considered to have higher probability to be correctly classified.

Algorithm 2 Self-learning SVM Method

Input: Training set $T = \{t_1, t_2, \dots, t_x\}$, T includes positive reviews and negative reviews; Unlabeled data set $U = \{u_1, u_2, \dots, u_y\}$;

Output: Sentiment classifier C ;

- 1: d is the number of selected reviews which are the most likely correctly classified reviews, $d=50$ in the experiment;
- 2: Use SVM to get the initial sentiment classifier C on the training data set T ;
- 3: while unlabeled data set is not empty do
- 4: Use sentiment classifier C to classify the unlabeled instances in U : get positive set P and the negative set N ;
- 5: If $|P| \geq d$, then select the most likely correctly classified d instances from P (the set is P_d) into T , $T = T \cup P_d$, $P = P - P_d$; otherwise put all the instances in the P into T , $T = T \cup P_d$;
- 6: If $|N| \geq d$, then select the most likely correctly classified d instances from N (the set is N_d) into T , $T = T \cup N_d$, $N = N - N_d$; otherwise put all the instances in the N into T , $T = T \cup N_d$;
- 7: Employ SVM to train a new sentiment classifier C on the current training data set T ;
- 8: end while;
- 9: return C .

5. EXPERIMENTAL RESULTS

Experimental data sets include three data sets which are phone, laptop, hotel reviews respectively. We randomly sample 800 reviews from each data set to form the test data sets. Each test set includes 400 positive reviews and 400 negative reviews. For each data set, we also sample 100, 200, 300, and 400 reviews for training respectively. The remaining reviews are unlabeled reviews. A training data set is a balanced training data set, for instance, 100 training instances include 50 positive instances and 50 negative instances. This work compares the proposed method, namely the Co-Training SVM method, with the Self-learning SVM method as Algorithm 1 shows. Accuracy is used to evaluate the proposed method.

$$\text{Accuracy} = \frac{\text{Number of correctly classified instances}}{\text{Number of total test instances}}$$

In Figure 3, we can see the proposed method can get higher sentiment classification accuracy than the Self-learning SVM method. The results on the phone data set show that the Two-view Co-training SVM method significantly improves with the increasing number of training instances, while the Self-learning SVM has the lowest accuracy when the number of the training data set is 200.

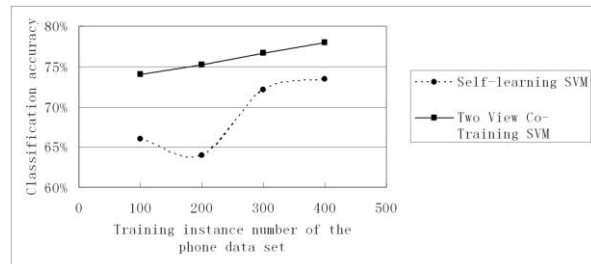


Fig 3: Sentiment classification accuracy on the phone data set

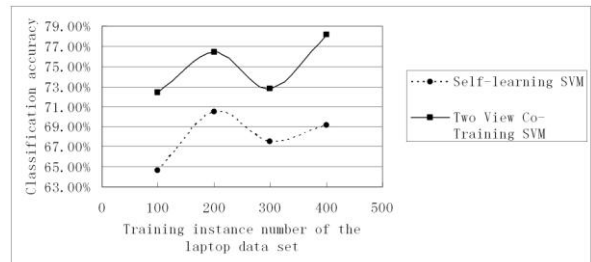


Fig 4: Sentiment classification accuracy on the laptop data set

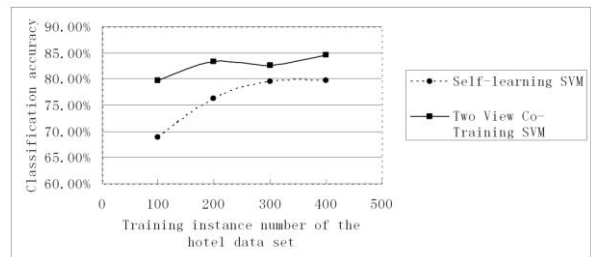


Fig 5: Sentiment classification accuracy on the hotel data set

⁶ <http://chasen.org/~taku/software/TinySVM/>

In Figure 4 and Figure 5, the Two-view Co-training SVM method outperforms the Self-learning SVM method in all different cases.

6. CONCLUSION

This paper proposes a method of semi-supervised sentiment classification method which exploits both subjective and objective views on co-training framework to do sentiment classification. The subjective view is constructed by the extracted opinion words from each consumer review. Non-opinion text features comprise of the objective view of each consumer review. Experimental results show the proposed method is effective, and it outperforms the Self-learning SVM method.

7. REFERENCES

- [1] A. Blum and T. Mitchell. "Combining labeled and unlabeled data with co-training", In Proceedings of the eleventh Annual Conference on Computational Learning Theory, COLT'98, pp.92–100, 1998.
- [2] V. Hatzivassiloglou and K.R. McKeown. "Predicting the semantic orientation of adjectives", In Proceedings of the eighth Conference on European chapter of the Association for Computational Linguistics, EACL '97, pp. 174–181, 1997.
- [3] P.D. Turney and M.L. Littman, "Measuring praise and criticism: Inference of semantic orientation from association", *ACM Trans. Inf. Syst.*, Vol. 21, No. 4, pp. 315–346, 2003.
- [4] P.D. Turney, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews", In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02, pp. 417–424, 2002.
- [5] B. Pang, L. Lee, and S. Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques", In Proceedings of the ACL-02 Conference on Empirical methods in Natural Language Processing, Vol. 10, EMNLP '02, pp. 79–86, 2002.
- [6] B. Pang and L. Lee. "Opinion mining and sentiment analysis", *Found. Trends Inf. Retr.*, Vol. No. (1-2), pp. 1–135, 2008.
- [7] R. Moraes, J.F. Valiati, and W.P.G Neto. "Document-level sentiment classification: An empirical comparison between SVM and ANN", *Expert Syst. Appl.*, Vol. 40, No. 2, pp. 621–633, 2013.
- [8] R. Xia, C. Zong, and S. Li. "Ensemble of feature sets and classification algorithms for sentiment classification", *Inf. Sci.*, Vol. 181, No. 6, pp. 1138–1152, 2011.
- [9] A. J. Blitzer, R. McDonald, and F. Pereira. "Domain adaptation with structural correspondence learning", In Proceedings of Conference on Empirical Methods in Natural Language Processing, pp. 120–128, 2006.
- [10] S.J. Pan, X. Ni, J. Sun, Q. Yang, and Z. Chen. "Cross-domain sentiment classification via spectral feature alignment", In Proceedings of the 19th International Conference on World Wide Web, WWW '10, pp. 751–760, 2010.
- [11] D. Bollegala, D. Weir, and J. Carroll. "Using multiple sources to construct a sentiment sensitive thesaurus for cross-domain sentiment classification", In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 132–141, 2011.
- [12] S. Gao and H. Li. "A cross-domain adaptation method for sentiment classification using probabilistic latent analysis", In Proceedings of the 20th ACM International Conference on Information and Knowledge Management, CIKM' 11, pp. 1047–1052, 2011.
- [13] L. Qiu, W. Zhang, C. Hu, and K. Zhao. "Selc: a self supervised model for sentiment classification", In Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM '09, pp. 929–936, 2009.
- [14] K. Sarvabhotla, P. Pingali, and V. Varma. "Sentiment classification: a lexical similarity based approach for extracting subjectivity in documents", *Inf. Retr.*, Vol. 14, No. 3, pp. 337–353, 2011.
- [15] S. Zhou, Q. Chen, and X. Wang. "Active deep networks for semi-supervised sentiment classification", In Proceedings of the 23rd International Conference on Computational Linguistics: Posters, COLING '10, pp.1515–1523, 2010.
- [16] S. Li, C. Huang, G. Zhou, and S. Y. M. Lee. "Employing personal/impersonal views in supervised and semi-supervised sentiment classification", In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL '10, pp. 414–423.
- [17] S. Li and J. Hao. "Spectral clustering-based semi-supervised sentiment classification", In Proceeding of 8th International Conference of Advanced Data Mining and Applications, pp. 271–283, 2012.
- [18] S. Li, Z. Guan, L. Tang, and Z. Chen. "Exploiting consumer reviews for product feature ranking", *J. Comput. Sci. Technol.*, Vol. 27, No.3, pp. 635–649, 2012.
- [19] C. Cortes and V. Vapnik, "Support-vector networks", *Machine Learning*, Vol. 20, No. 3, pp. 273–297, 1995.