Machine Learning and Lexicon based Methods for Sentiment Classification: A Survey [1]

Hailong Zhang, Wenyan Gan, Bo Jiang
Institute of Command Information Systems, PLA University of Science and Technology,
Nanjing, China
e-mail: purplehero3@qq.com, wenyangan@163.com

***Abstract*—Sentiment classification is an important subject in text mining research, which concerns the application of automatic methods for predicting the orientation of sentiment present on text documents, with many applications on a number of areas including recommender and advertising systems, customer intelligence and information retrieval. In this paper, we provide a survey and comparative study of existing techniques for opinion mining including machine learning and lexicon-based approaches, together with evaluation metrics. Also cross-domain and cross-lingual approaches are explored. Experimental results show that supervised machine learning methods, such as SVM and naive Bayes, have higher precision, while lexicon-based methods are also very competitive because they require few effort in human-labeled document and isn't sensitive to the quantity and quality of the training dataset.**

***Keywords*—Sentiment classification; Performance; Lexicon; Machine Learning;Cross-domain; Cross-lingual；Deep learning**

I. INTRODUCTION

With the emergence and rapid development of Web 2.0, more and more people begin to express their feelings, opinion and attitude over Internet, which increase the amount of user-generated reviews containing rich opinion and sentiment information. Sentiment analysis (also known as opinion mining), as a technique to automatic detection of opinions embodied in text, is becoming a hotspot in many research fields, including natural language processing(NLP), data mining (DM) and information retrieval (IR), with a number of applications including recommender and advertising systems, product feedback analysis and customer decision making.

However, finding opinion sources and monitoring them on the Web can still be a formidable task because there are a large number of diverse sources, and each source may also have a huge volume of *opinionated text* (text with opinions or sentiments). Two type of classification techniques have been used in document-level sentiment classification: manual or human sentiment classification and automated sentiment classification.

Considering that sentiment words and phrases are the main indicators of sentiment classification, early work in sentiment analysis mainly involved the use of knowledge-based or lexicon-based approaches, which make use of publicly available lexicon resources, e.g. WordNet or SentiWordNet, and classify the sentiment of texts based on dictionaries defining the sentiment-polarity of words. The main problem is that the lexicon-based approaches do not adapt well to different domains or different languages. A word may express positive emotion in one domain but negative emotion in another domain. So the classical lexicon-based approach should be improved to find sentiment words with domain specific orientation.

Recently, there have been some studies that take different machine learning approaches, and build text classifiers, such as decision tree, naive Bayes, and Support Vector Machines, etc. But this kind of techniques require much effort in human annotation of documents, usually is sensitive to the quantity and quality of the training dataset and may fail when training data are biased or insufficient.

In this paper, we provide a survey and comparative study of existing techniques for opinion mining including machine learning and lexicon-based approaches, together with evaluation metrics. Experimental results show that supervised machine learning methods, such as SVM and naive Bayes, have higher precision, while lexicon-based methods are also very competitive because they require few effort in human-labeled document and isn't sensitive to the quantity and quality of the training dataset. With the imbalance of the sentiment resources in different languages and domains, cross-domain and cross-lingual sentiment classification methods are also put forward to solve this problem.

II. SENTIMENT CLASSIFICATION BASED ON MACHINE LEARNING

A. Machine Learning Method

Machine learning method trains a text classifier on a human-labeled training dataset. Sentiment classification based on machine learning can be formulated as a supervised learning problem. One of the earliest works includes Pang et al.[19], which used the standard bag of feature framework and a variety of machine learning techniques to solve sentiment classification problem. They test several features to fined optimal feature set: unigrams, bigrams, adjective and position of words were used as features, and found that the best performance was achieved when the unigrams were used in SVM classifier. In the later work, Pang and Lee[20] reported improvement by adding a preprocessing filter to remove objective sentences which allowed the classifier to focus only on subjective sentences, raising the accuracy from 82.9% to 86.4% in a movie reviews dataset. Cui et al.[9] also argued that discriminative classifiers such as SVMs were more appropriate

[1] This work is supported by the National Natural Science Foundation of China under Grant No. 60974086.

978-1-4799-5727-9/14 $31.00 © 2014 IEEE
DOI 10.1109/WISA.2014.55

for sentiment classification than generative models because they can differentiate mixed sentiment better. However, when the set of training data is small, a naive Bayes classifier might be more appropriate since SVM must be exposed to a large set of data in order to build a high-quality classifier.

Considering that supervised machine learning techniques require a large corpus of training data and its performance depends on a good match between the training and testing data with respect to the domain, topic and time-period, some semi-supervised or weakly-supervised approaches [21-23] have been used for sentiment classification on many mediums, whose performance may be less dependent on the domain, topic and time-period represented by the testing data. Liu, S. et al.[7] formally proposed an adaptive multiclass SVM model which transfers an initial common sentiment classifier to a topic adaptive one. That makes sentiment classification adaptive to diverse topics without sufficient labeled data. Deep learning methods are new hotspot [29] in machine learning research which recently is used to solve large-scale unsupervised sentiment classification, in which each layer of a deep neural network architecture represents features at a different level of abstraction, defined as a composition of lower-level features[10-12,30]. As an unsupervised learning method, sentiment classification based on deep learning requires more attention.

B. Feature Selection For Sentiment Analysis

Feature selection is also a critical task in sentiment analysis and effectively selected representative features from subjective text can improve sentiment based classification. Pang et al.[19] classify movie reviews into two classes, positive and negative. It was shown that using unigrams as features in classification performed well with either naive Bayesian or SVM. Pang and Lee [1] make a comprehensive survey of features used. Subsequent researches used many more kinds of features and techniques in learning.

Sharma A. & Dey S.[5] explored the applicability of five commonly used feature selection methods in data mining research (DF, IG, GR, CHI and Relief-F) and seven machine learning based classification techniques (Naive Bayes, Support Vector Machine, Maximum Entropy, Decision Tree, K-Nearest Neighbor, Winnow, Adaboost) for sentiment analysis on online movie reviews dataset. The work demonstrates that feature selection does improve the performance of sentiment based classification, but it depends on the method adopted and the number of feature selected. The experimental results show that Gain Ratio gives the best performance for sentimental feature selection, and SVM performs better than other techniques for sentiment based classification. Since the domain-specific nature of sentiment classification led the task more problematic, considering more contextual-information such as topic or domain is essential[15].

III. Sentiment Classification based on Opinion Lexicon

Opinion lexicon methods adopt a lexicon to perform sentiment analysis, by counting and weighting sentiment-related words that have been evaluated and tagged. To collect the opinion word list, three main opinion approaches have been investigated: the manual approach, the dictionary-based approach, and the corpus-based approach. The manual approach is very time consuming and thus it is not usually used alone, but combined with automated approaches as the final check because automated methods make mistakes. In the following, two automatic methods will be discussed.

Corpus-based methods use a seed set of sentient words with known polarity and exploit syntactic patterns of co-occurrence patterns to identify new sentiment words and their polarity in a large corpus. For example, Turkey determined whether words are positive or negative and how strong the evaluation is by computing the words' point-wise mutual information (PMI) for their co-occurrence with the word's sentiment orientation. This method scanned through a review looking for phrases that match certain part of speech patterns (adjective and adverbs), and then added up all sentiment orientation to compute the orientation of a document. He achieves 74% accuracy classifying a corpus of product reviews.

Dictionary-based methods exploit available lexicographical resources like WordNet or HowNet. The main strategy in these methods is to collect an initial seed set of sentiment words and their orientation manually, and then searching in a dictionary to find their synonyms and antonyms to expand this set. The new seed set are used iteratively to generate new sentiment words. Taboada, M. et al.[17] present a lexicon-based approach which uses Semantic Orientation CALculator (SO-CAL) to extracting sentiment from text. SO-CAL's performance is consistent across domains and on completely unseen data.

Using corpus-based approach alone to identify all opinion words, is not as effective as the dictionary-based approach because it is hard to prepare a huge corpus to cover all words. However, this approach has a major advantage it can help to find domain-specific opinion words and their orientations if a corpus from only the specific domain is used in the discovery process. Ding et al. [18] further explore the idea of intra-sentential and inter-sentential sentiment consistency. Instead of finding domain dependent opinion words, they showed that the same word might have different orientations in different contexts even in the same domain.

IV. Cross-domain and cross-lingual sentiment classification

In addition, the imbalance of the sentiment resources in different languages and domains also challenges sentiment classification task. Therefore sentiment classification based on cross-domain and cross-lingual methods are put forward to solve this problem.

A. Cross-Lingual Sentiment Classification

The task of sentiment classification relies heavily on sentiment resources, including annotated lexicons and corpus. However, the sentiment resources in different languages are imbalanced. For example, the lack of Chinese sentiment corpora limits the research progress on Chinese sentiment classification. However, there are many freely available English sentiment corpora on the Web. We focus on the problem of cross-lingual sentiment classification, which leverages an available English corpus for Chinese sentiment classification by using the English corpus as training data. Machine translation services are used for eliminating the language gap between the training set and test set, and English features and Chinese features are considered as two independent views of the classification problem. Wan, X.[4] propose a co-training approach to make use of unlabeled Chinese data. And that shows the effectiveness of the proposed approach, which can outperform the standard inductive classifiers and the transductive classifiers. A few novel models have been proposed to address the problem, e.g. the EM-based algorithm[16], the information bottleneck approach, the multilingual domain models, etc.

Meng et al.[13] propose a generative cross-lingual mixture model (CLMM) to leverage unlabeled bilingual parallel data. By fitting parameters to maximize the likelihood of the bilingual parallel data, it learns previously unseen sentiment words from the large bilingual parallel data and improves vocabulary coverage significantly. For the classification of Web forum opinions in multiple languages, Abbasi, A. et al.[28] propose a entropy weighted genetic algorithm (EWGA) method which is a hybridized genetic algorithm that incorporates the information-gain heuristic for feature selection. Wan, X.[3] conduct a comparative study to explore the challenges of cross-lingual sentiment classification. Different schemes for cross-lingual sentiment classification based on two dimensions have been compared empirically. He proposed to combine the different individual schemes into an ensemble. The proposed method gets a high effectiveness.

B. *Cross-Domain Sentiment Classification*

Since the domain-specific nature of sentiment classification led the task more problematic, considering more contextual-information such as topic or domain is essential. In the problem of cross-domain text classification, the labeled and unlabeled data come from different domains, and their underlying distributions are often different from each other, which violates the basic assumption of traditional classification learning.

Many semi-supervised learning algorithms have been developed for addressing the cross-domain text classification problem by transferring knowledge across domains, including Transductive SVM [27], EM [16], EM-based naive Bayes classifier[24], Topic-bridged PLSA [14], Co-Clustering based classification[25], two-stage approach[26].

Li, S. et al.[6] suggest to perform active learning for cross-domain sentiment classification by actively selecting a small amount of labeled data in the target domain. In their work, two individual classifiers are trained, i.e., the source and target classifiers with the labeled data from the source and target respectively. Then, the two classifiers are employed to select informative samples with the selection strategy of Query By Committee (QBC). And the two classifiers are combined to make the classification decision. Importantly, the two classifiers are trained by fully exploiting the unlabeled data in the target domain with the label propagation (LP) algorithm.

Moreover, Bollegala, D.et al.[8] create a sentiment sensitive distributional thesaurus using labeled data for the source domains and unlabeled data for both source and target domains. Sentiment sensitivity is achieved in the thesaurus by incorporating document level sentiment labels in the context vectors used as the basis for measuring the distributional similarity between words. Pan S J. et al.[2] propose a spectral feature alignment (SFA) algorithm to align domain-specific words from different domains into unified clusters, with the help of domain-independent words as a bridge. SFA can discover a robust representation for cross-domain data by fully exploiting the relationship between the domain-specific and domain-independent words via simultaneously co-clustering them in a common latent space.

V. EVALUATION OF SENTIMENT CLASSIFICATION

The performance of sentiment classification can be evaluated by using four indexes calculated as the following equations: Accuracy, Precision, Recall and F1 score.

$$\mathbf{Accuracy = (TP+TN)/(TP+TN+FP+FN)} \quad (1)$$

$$\mathbf{Precision = TP/(TP+FP)} \quad (2)$$

$$\mathbf{Recall = TP/(TP+FN)} \quad (3)$$

$$\mathbf{F1 = (2 \times Precision \times Recall)/(Precision+Recall)} \quad (4)$$

In which TP, FN, FP and TN refer respectively to the number of true positive instances, the number of false negative instances, the number of false positive instances and the number of true negative instances, as defined in the table 1.

Table1: Confusion Matrix

	Predicted Positives	Predicted Negatives
Actual Positives	TP	FN
Actual Negative	FP	TN

According to the above performance measures, we provide a comparative study of existing techniques for opinion mining, including machine learning, lexicon-based approaches, cross-domain and cross-lingual approaches, etc., as shown in Table 2.

Table2. Performance comparison of sentiment classification methods

	Method	Data Set	Acc.	Author
Machine learning	SVM	Movie reviews	86.40%	Pang ,Lee[20]
	CoTraining SVM	Twitter	82.52%	Liu[7]
	Deep learning	Stanford Sentiment Treebank	80.70%	Richard [12]
Lexicon based	Corpus	Product reviews	74.00%	Turkey
	Dictionary	Amazon’s Mechanical Turk	---	Taboada[17]
Cross-lingual	Ensemble	Amazon	81.00%	Wan, X.[3]
	Co-Train	Amazon,IT168	81.30%	Wan, X.[4]
	EWGA	IMDb movie review	>90%	Abbasi, A.[28]
	CLMM	MPQA,NTCIR,ISI	83.02%	Meng[13]
Cross-domain	Active Learning	Book, DVD, Electronics, Kitchen	80% (average)	Li,S.[6]
	Thesaurus			Bollegala [8]
	SFA			Pan S J[2]

Further, we build a simple lexicon-based sentiment classification based on SentiWordNet and evaluate it on the SFU Review Corpus in order to compare its accuracy with the baseline learning-based methods, SVM and naive Bayes. As shown in Table 3, SentiWrodNet-based approach seems to be competitive, getting its best results using only adjectives, and also has the additional advantage that it does not rely on the quality or amount of training data. According to the above results, we think that it may be interested in building an integrated framework combining opinion lexicon and machine learning to perform unsupervised or semi-supervised large-scale sentiment classification with improved performance and adaptation to different domains and languages.

Table3 Performance comparison of sentiWordNet and learning-based methods

	TP	FP	FN	TN	F1	Accuracy
Senti-WordNet	148	91	52	109	63.91%	64.25%
NB	156	81	44	119	68.48%	68.75%
SVM	135	51	65	149	70.96%	71.00%

VI. CONCLUSION AND FUTURE WORKS

In this paper, we provide a survey and comparative study of existing techniques for opinion mining including machine learning and lexicon-based approaches, together with cross-domain and cross-lingual methods and some evaluation metrics. Research results show that machine learning methods, such as SVM and naive Bayes have the highest accuracy and can be regarded as the baseline learning methods, while lexicon-based methods are very competitive, which require few effort in human-labeled document. And deep learning methods are also studied to solve this problem. In the future work, we will focus on the study on combining machine learning method into opinion lexicon method in order to improve the accuracy of sentiment classification and adaptive capacity to different domains and language.

REFERENCES

[1] Pang B and Lee L, Opinion mining and sentiment analysis. Foundations and Trends in IR. 2008. 1-135.

[2] Pan S J, Ni X, Sun J T, et al. Cross-domain sentiment classification via spectral feature alignment[C]//Proceedings of the 19th international conference on World wide web. ACM, 2010: 751-760.

[3] Wan, X.. A Comparative Study of Cross-Lingual Sentiment Classification. In Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01 (pp. 24-31).IEEE Computer Society.2012

[4] Wan X. Co-training for cross-lingual sentiment classification[C]// Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1. Association for Computational Linguistics, 2009: 235-243.

[5] Sharma, A., & Dey, S.. A comparative study of feature selection and machine learning techniques for sentiment analysis. In Proceedings of the 2012 ACM Research in Applied Computation Symposium (pp. 1-7). ACM,2012

[6] Li, S., Xue, Y., Wang, Z., & Zhou, G.. Active learning for cross-domain sentiment classification. In Proceedings of the Twenty-Third international joint conference on Artificial Intelligence (pp. 2127-2133). AAAI Press,2013

[7] Liu, S., Li, F., Li, F., Cheng, X., & Shen, H.. Adaptive co-training SVM for sentiment classification on tweets. In Proceedings of the 22nd ACM international conference on Conference on information & knowledge management (pp. 2079-2088). ACM,2013.

[8] Bollegala, D., Weir, D., & Carroll, J.. Cross-Domain Sentiment Classification using a Sentiment Sensitive Thesaurus. Knowledge and Data Engineering, IEEE Transactions on, 25(8), 1719-1731,2013.

[9] H. Cui, V. Mittal, and M. Datar. Comparative Experiments on Sentiment Classification for Online Product Reviews. In Proceedings of AAAI-06, pp.1265-1270, 2006.

[10] Glorot, Xavier, Antoine Bordes, and Yoshua Bengio. "Domain adaptation for large-scale sentiment classification: A deep learning approach." Proceedings of the 28th International Conference on Machine Learning (ICML-11). 2011.

[11] Bespalov, Dmitriy, et al. "Sentiment classification based on supervised latent n-gram analysis." Proceedings of the 20th ACM international conference on Information and knowledge management. ACM, 2011.

[12] Socher, Richard, et al. "Recursive deep models for semantic compositionality over a sentiment treebank." Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). 2013.

[13] Meng, Xinfan, et al. "Cross-lingual mixture model for sentiment classification." Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. Association for Computational Linguistics, 2012.

[14] G.-R. Xue, W. Dai, Q. Yang, Y. Yu. Topic-bridged PLSA for cross-domain text classification. In Proceedings of SIGIR-08.

[15] Choi, Y., Kim, Y., & Myaeng, S. H.. Domain-specific sentiment analysis using contextual feature generation. In Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion (pp. 37-44). ACM, 2009.

[16] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell. Text Classification from Labeled and Unlabeled Documents using EM. Machine Learning, 39(2-3):103 – 134, 2000.

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

[18] X. Ding, B. Liu, and P. S. Yu, A holistic lexicon-based approach to opinion mining, Proceedings of the Conference on Web Search and Web Data Mining (WSDM), Stanford, CA, 2008.

[19] Pang, B., Lee, L., and Vaithyanathan. Sentiment classification using machine learning techniques[A]. Conference on Empirical Methods in Natural Language Processing [C](EMNLP-02). 2002, 79-86.

[20] Pang, B., and Lee, L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts[A]. 42nd Meeting of the Association for Computational Linguistics[C] (ACL-04). 2004, 271-278.

[21] Read J, Carroll J. Weakly supervised techniques for domain-independent sentiment classification[C]//Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion. ACM, 2009: 45-52.

[22] S. Li, Z. Wang, G. Zhou, and S. Y. M. Lee.Semi-supervised learning for imbalanced sentiment classification. In Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume Three, pages 1826 – 1831. AAAI Press, 2011.

[23] N. Yu and S. K ¨ ubler. Filling the gap: Semi-supervised learning for opinion detection across domains. In Proceedings of the Fifteenth Conference on Computational Natural Language Learning, pages 200 – 209. Association for Computational Linguistics, 2011.

[24] L. Rigutini, M. Maggini and B. Liu. An EM based training algorithm for cross-language text categorization. In Proceedings of WI-05.

[25] W. Dai, G.-R. Xue, Q. Yang, Y. Yu. Co-clustering based classification for out-of-domain documents. In Proceedings of KDD-07.

[26] J. Jiang and C. Zhai. A two-stage approach to domain adaptation for statistical classifiers. In Proceedings of CIKM-07.

[27] T. Joachims. Transductive inference for text classification using support vector machines. In Proceedings of ICML-99.

[28] Abbasi A, Chen H, Salem A. Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums[J]. ACM Transactions on Information Systems (TOIS), 2008, 26(3): 12.

[29] HINTON, Geoffrey E.; OSINDERO, Simon; TEH, Yee-Whye. A fast learning algorithm for deep belief nets. Neural computation, 2006, 18.7: 1527-1554.

[30] Zhou, Shusen, Qingcai Chen, and Xiaolong Wang. "Active deep networks for semi-supervised sentiment classification." Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Association for Computational Linguistics, 2010.