Feature Subsumption for Sentiment Classification in Multiple Languages

Zhongwu Zhai, Hua Xu, Jun Li, and Peifa Jia

State Key Laboratory on Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology
CS&T Department, Tsinghua University, Beijing 100084, China P.R.
{zhaizhongwu,junli.cn}@gmail.com

Abstract. An open problem in machine learning-based sentiment classification is how to extract complex features that outperform simple features; figuring out which types of features are most valuable is another. Most of the studies focus primarily on character or word *Ngrams* features, but *substring-group* features have never been considered in sentiment classification area before. In this study, the substring-group features are *extracted* and *selected* for sentiment classification by means of *transductive* learning-based algorithm. To demonstrate generality, experiments have been conducted on *three* open datasets in *three different languages*: Chinese, English and Spanish. The experimental results show that the proposed algorithm's performance is usually superior to the best performance in related work, and the proposed feature subsumption algorithm for sentiment classification is *multilingual.* Compared to the *inductive* learning-based algorithm, the experimental results also illustrate that the *transductive* learning-based algorithm can significantly improve the performance of sentiment classification. As for term weighting, the experiments show that the "tfidf-c" outperforms all other term weighting approaches in the proposed algorithm.

Keywords: Sentiment, Transductive, Substring-group, Multilingual.

1 Introduction

With the growing availability and popularity of online user-generated information, including reviews, forum discussions, and blogs, sentiment analysis and opinion mining ("sentiment analysis" and "opinion mining" denote the same field of study [1]) have become one of the key technologies for handling and analyzing the text data from internet. One of the most widely-studied sub-problems of opinion mining is sentiment classification, which classifies evaluative documents, sentences or words as positive or negative (in some cases, the neutral class is used as well) [2] to help people automatically identify the viewpoints underlying the online user-generated information [3]. Since sentiment classification concerns the opinion expressed in a text rather than its topic, it challenges data-driven methods and resists conventional text classification techniques [4].

M.J. Zaki et al. (Eds.): PAKDD 2010, Part II, LNAI 6119, pp. 261–271, 2010.

Up to this date, machine learning-based methods have been commonly adopted for sentiment classification due to their outstanding performance [3, 4]. An open problem in machine learning-based sentiment classification is how to extract complex features that outperform simple features; figuring out which types of features are most valuable is another [5]. Most of the existing research focus on simple features, including single words [6], character Ngrams [7, 8], word Ngrams [3, 4, 8], phrases [5] and the combination of above features, but *substring-group* features have never been considered in sentiment classification. In fact, the substring-group features based classification approaches have at least the following potential advantages [9], allowing sub-word features and super-word features to be exploited automatically. With such approaches, the messy and rather artificial problem of defining word boundaries in some Asian languages can be avoided, and non-alphabetical features can be taken into account. Furthermore, different types of documents can be dealt with in a uniform way.

In this study, the substring-group features are extracted and selected for sentiment classification by means of the *transductive learning*-based algorithm. **Firstly**, the substring-group features are extracted from the suffix tree constructed by the training and unlabeled test documents, based on the *transductive learning* theory [10]. Since the *substring-groups* include several continuous words or even sentences, the substring-group features facilitate the incorporation of word sequence information to sentiment classification. Also, since the suffix tree is constructed by both training documents and unlabeled test documents, the *structural information of unlabeled test documents* is incorporated to feature extraction. **Secondly**, the extracted substring-group features are further selected to eliminate the redundancy among them. At last, SVM is adopted to classify the unlabeled test documents based on the selected features. Experiments have been conducted on three open datasets in three different languages, Chinese, English and Spanish, and the experimental results demonstrate the effectiveness of the proposed algorithm.

The rest of this paper is organized as follows. Section 2 reviews the learning paradigms and the related work. The proposed algorithm is described in detail in Section 3. The experimental setup is illustrated in Section 4 and the results are given and analyzed in Section 5. Finally, this paper is summarized in Section 6.

2 Learning Paradigms and Related Work

Learning paradigms: Given an example x and a class label y, the standard statistical classification task is to assign a probability, $Pr(y|x)$, to x of belonging to class y. In sentiment classification, the labels are $Y \in \{$*'positive'*, *'negative'*$\}$. The data for the sentiment classification task consists of two disjoint subsets: the training set $(X_{train}, Y_{train}) = \{(x_1, y_1), \cdots, (x_N, y_N)\}$, available to the model for its training, and the test set $X_{test} = (\mathrm{x}_1, \cdots, \mathrm{x}_M)$, upon which we want to leverage the trained classifier to make predictions.

In the paradigm of ***inductive learning***, (X_{train}, Y_{train}) are known, while both X_{test} and Y_{test} are completely hidden during training time. In the case of semi-supervised inductive learning [10-12], the learner is also provided with auxiliary unlabeled data $X_{auxiliary}$, that is not part of the test set. Another setting that is closely related to semi-supervised learning is ***transductive learning*** [10, 13, 14], in which X_{test} (but, importantly, not Y_{test}), is known at training time. One can think of transductive learning as a special case of semi-supervised learning in which $X_{auxiliary} = X_{test}$.

Related work: Sentiment classification can be performed on word level, sentence level and document level. In this paper, we focus on document sentiment classification. Previous studies for sentiment classification on document level can be generally classified into two categories, unsupervised approaches and supervised approaches.

The unsupervised approaches focus on identifying semantic orientation of individual words or phrases, and then classifying each document in terms of the number of these words or phrases contained in each document. Turney determines semantic orientation by phrase Pointwise Mutual Information (PMI) based on predefined seed words [15] and rates reviews as thumbs up or down [16]. Kim and Hovy [17] build three models to assign a sentiment category to a given sentence by combining the individual sentiments of sentiment-bearing words. Liu et al. classify customer reviews using a holistic lexicon [18, 19]. Kennedy and Inkpen determine the sentiment of customer reviews by counting positive and negative terms and taking into account contextual valence shifters, such as negations and intensifiers [20]. Devitt and Ahmad explore a computable metric of positive or negative polarity in financial news text [21]. Wan uses bilingual knowledge and ensemble techniques for unsupervised Chinese sentiment analysis [22].

The supervised approaches focus on training a sentiment classifier using labeled corpus. Since the work of Pang et al. [4], various classification models and linguistic features have been proposed. Dave et al. use machine learning based methods to classify reviews on several kinds of products [23]. Pang and Lee report 86.4% accuracy rate of sentiment classification of movie reviews by using word unigrams features for SVMs [3]. Mullen and Collier also employ SVMs to bring together diverse sources of potentially pertinent information, including several favorability measures for phrases and adjectives and knowledge of the topic of the text [24]. Most recently, Li and Sun compare the performance of four machine learning methods for sentiment classification of Chinese reviews using Ngrams features [8]. Blitzer et al. investigate domain adaptation for sentiment classifier [25]. Songbo et al. combine learn-based and lexicon-based techniques for sentiment detection without using labeled examples [26].

To the best of our knowledge, though substring-group features have been used for topic, authorship and genre classification [9], they have not yet been considered in sentiment classification. Moreover, the structural information of *unlabeled* test documents is used in this paper by *transductive learning*, which has not been studied in any related work of sentiment classification. Furthermore, the synergetic effect of feature extracting and feature selecting has been reflected in this study.

3 The Proposed Algorithm

According to the conclusions from the learning paradigms in subsection **2.1**, not only the training documents, but also the unlabeled test documents can be used at training time by the *transductive* learning-based algorithm [10]. The proposed algorithm takes the training documents (*both text and class labels*) and the unlabeled test documents (*only text*) as input, and outputs the predicted classifications of the unlabeled test documents. The framework of the proposed algorithm is shown in Figure **1**, including four stages: *substring-group feature extracting*, term weighting, *feature selecting* and classifying.

3.1 Substring-Group Feature Extracting

The unique substring-group features are extracted by the following steps.

Step (***a***) aims to construct a suffix tree using all the strings of both *training* documents and *unlabeled test* documents. The suffix tree is constructed by Ukkonen's algorithm with *O(n)* time complexity, where n is the number of characters in the text corpus [27]. This step shows the incorporation of *transductive learning*.

In step (***b***), the key-nodes are extracted from the constructed suffix tree. For an *m-character* text corpus, the constructed suffix tree has *m* leaf-nodes and at most *m-1* internal nodes [28]. The text corpus's length *m* is usually a very large number, so it's necessary to extract the key-nodes from the (*2m-1*) nodes. The key-node extracting criteria proposed in [9] is used in this paper, and the recommended values are adopted: L=20, H=8000, B=8, P=0.8 and Q=0.8. The meanings of the parameters are listed in Table 1.

In step (***c***), every suffix of each document is matched with the suffix tree, and all the IDs of the matched key-nodes are taken as the content of the corresponding document.

In step (*d*), from the ***training*** part of the converted documents, all the unique key-node IDs are extracted as features for sentiment classification. This step guarantees that the evaluation of the following experiments is *open test*.

In step (***e***), all the converted documents are translated into corresponding vectors using the unique feature table produced by step (*d*).

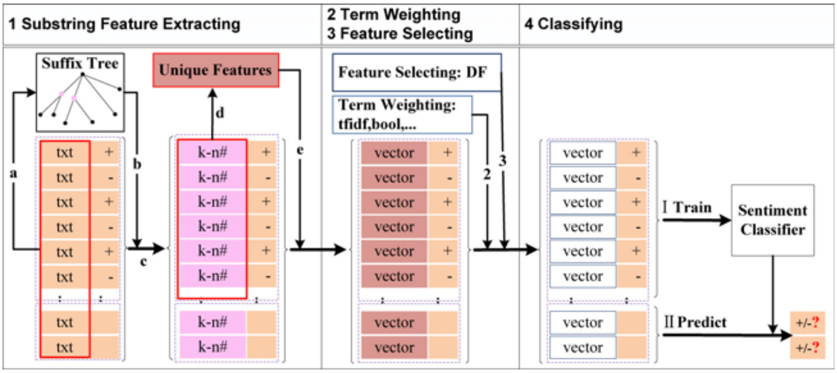

Fig. 1. The framework of the proposed algorithm

According to the definition of suffix tree [28], each node of the suffix tree represents a substring-group of the text corpus. Therefore, the extracted key-node IDs are also called substring-group features in this paper. Moreover, the time complexity of the computing steps a, b and c is linear, which has been proved in [9].

Table 1. The key-nodes extracting parameters

L	The minimum frequency. A node is not extracted, if it has less than L leaf-nodes in the suffix tree.
H	The maximum frequency. A node is not extracted, if it has more than H leaf-nodes in the suffix tree.
B	The minimum number of children. A node is not extracted, if it has less than B children.
P	The maximum parent-child conditional probability. A node u is not extracted, if the probability Pr(v|u) = freq(v)/freq(u) ≥ P, where u is the parent node of v.
Q	The maximum suffix-link conditional probability. A node s(v) is not extracted, if the probability Pr(v|s(v)) = freq(v)/freq(s(v)) ≥ Q, where the suffix-link of v points to s(v).

3.2 Term Weighting

Term frequency has traditionally been used in the standard text classification, but Pang et al. [4] obtained better performance by using presence rather than frequency. Consequently, both term presence ("bool", "three") and term frequency ("tf" and "tfidf-c") are used in this paper. The "tfidf-c" is the variants of standard "tfidf", and it is widely used in text classification [29, 30]. The four adopted term weighting approaches are defined as formulas 1, 2, 3, and 4.

$$bool: \quad \begin{cases} 1 & if\ tf(t_k, d_j) > 0 \\ 0 & if\ tf(t_k, d_j) = 0 \end{cases} \tag{1}$$

$$three: \quad \begin{cases} 2 & if\ tf(t_k, d_j) > 1 \\ 1 & if\ tf(t_k, d_j) = 1 \\ 0 & if\ tf(t_k, d_j) = 0 \end{cases} \tag{2}$$

$$tf: \quad tf(t_k, d_j) \tag{3}$$

$$tfidf\text{-}c: \quad \frac{tf(t_k, d_j) \times log\dfrac{N}{df(t_k)}}{\sqrt{\sum_{t \epsilon d_j}\left(tf(t_k, d_j) \times log\dfrac{N}{df(t_k)}\right)^2}} \tag{4}$$

Here, t_k denotes a distinct term corresponding to a single feature; $tf(t_k, d_j)$ represents the number of times term t_k occurs in the document d_j; $df(t_k)$ is the number of documents the term t_k occurs in; N is the total number of training documents.

3.3 Feature Selecting

Document frequency (DF) is the number of documents in which a term occurs. It is the simplest criterion for feature selection and can be easily scaled to a large dataset with linear computation complexity. It is a simple but effective feature selection method for text categorization [31]. In this study, DF is used to pick out the discriminating substring-group features for training and classification.

For DF (Document Frequency) calculation, we compute the document frequency for each feature in the training corpus and then select the top N features with the highest scores. The basic assumption is that the rare features are either non-informative for class prediction, or not influential in global performance.

3.4 Classifying

In this step, the sentiment classifier is trained by machine learning algorithms to predict the classifications of the unlabeled test documents. Due to SVMs' outstanding performance [3, 4, 6, 8, 24, 32], SVMs are adopted in this paper. The SVM^{light} package is used for training and testing with default parameters.

4 Experimental Setup

4.1 Datasets

The proposed algorithm has been tested on three open datasets in three different languages: *Chinese*, *English* and *Spanish*. Table 2 gives a short summary of these open datasets.

Table 2. The summary of the open datasets

Language	**Positive**	**Negative**	**n-fold CV**	**Encoding**
Chinese_16000[1]	8000	8000	4	GB2312
English_1400[2]	700	700	3	ASCII
Spanish_400[3]	200	200	3	ISO-8859-2

These 160,000 *Chinese* hotel reviews were crawled from the website `http://www.ctrip.com/`, which is one of the most well-known websites in China for hotel and flight reservation. The "*English*_1400" is most commonly used for sentiment classification in English. The *Spanish* corpus is a collection of 400 reviews on cars, hotels, washing machines, books, cell phones, music, computers, and movies. Each category contains 50 positive and 50 negative reviews, defined as positive or negative based on the number of stars given by the reviewers.

In order to compare the results from the related works on these open datasets, *4-fold*, *3-fold* and *3-fold* **cross validation** are used respectively in the following experiments.

4.2 Evaluation Metrics

To evaluate the performance of the proposed algorithm for sentiment classification, we adopted traditional evaluation metric *accuracy* that is generally used in text categorization [30]. In addition, *microF1* and *macroPrecision* are also computed to compare with the related work.

[1] `http://nlp.csai.tsinghua.edu.cn/~lj/pmwiki/index.php?n=Main.DataSet`

[2] `http://www.cs.cornell.edu/people/pabo/movie-review-data/mix20_rand700_tokens.zip`

[3] `http://www.sfu.ca/~mtaboada/research/SFU_Review_Corpus.html`

5 Experimental Results

5.1 Comparisons

To compare to the algorithms in related work, the best performances of the existing typical methods on each dataset are listed in Table 3, respectively. The column "#Features" is the number of features when the best performance is achieved.

Table 3. Comparisons with the best performance of the existing typical methods

Language	Techniques	Best Performance(%)	#Features
Chinese (16,000)	SVM(word bigrams, tfidf-c) [8]	91.2^{microF1}	251,289
	SVM(character bigrams, tfidf-c) [8]	91.6^{microF1}	128,049
	SVM(*key substring-groups* + DF, tfidf-c)	$\mathbf{94.0}^{\text{microF1}}$	41,454
English (1,400)	SVM(character unigrams, bool) [4]	82.9^{accuracy}	16,165
	SVM(*key substring-groups* + DF, tfidf-c)	$\mathbf{84.3}^{\text{accuracy}}$	28,726
Spanish (400)	No existing work has used this corpus yet.		
	SVM(*key substring-groups* + DF, tfidf-c)	$\mathbf{78.7}^{\text{accuracy}}$	2,519

As illustrated in Table 3, although the proposed algorithm (shown in gray background) **does** not use any preprocessing steps, such as *word segmentation* and stemming, it outperforms the character or word Ngrams based methods on three different language datasets. *Note that all these datasets are processed by the proposed algorithm in a* ***uniform*** *way rather than different* ***language-specific*** *ways.* Another observation is that the number of features (#Features) used in the proposed algorithm is larger than most other algorithms, which indicates that the promising performance of the proposed algorithm is at the cost of high feature dimension.

5.2 Multilingual Characteristics

Since the proposed algorithm treats the input documents as character sequences regardless of their syntax or semantic structures, no word segmentation technology is needed. Consequently, the proposed algorithm can deal with any language in any encoding, which has been demonstrated by the experiments in Table **3**.

Furthermore, the proposed algorithm is capable of handling text corpus containing both English and Chinese words at the same time. We conduct an experiment on the mixed-language dataset, including the "English_1400" corpus and 1,400 Chinese reviews (700 pos + 700 neg) randomly selected from the "Chinese_16000" corpus. Three-fold cross validation is adopted. The experimental results are shown in Figure 2. As is shown in Figure 2, the proposed algorithm achieves promising performance (shown in dark blue curve) on the mixed-language dataset, which is even better than the performance obtained by using only the English corpus.

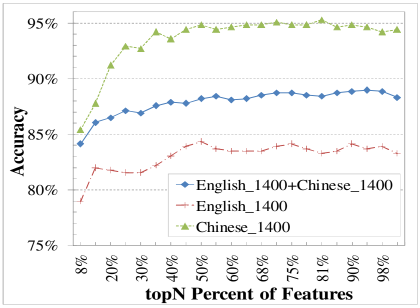

Fig. 2. The experiment on the mixed-language dataset (DF+"tfidf-c")

5.3 Feature Frequency vs. Feature Presence

The performance of sentiment classifier is highly affected by the text representation. To show the impact of the term weighting approaches, we conduct a series of experiments.

As demonstrated in Figure 3, different term weighting approaches lead to different classifying performances. Among all the term weighting methods, the "tfidf-c" outperforms all other approaches on the three open datasets in different languages, while the "tf" performs the worst. The "bool" always achieves better performance than the "three".

This observation agrees with Pang's finding: the better performance is achieved by accounting only for feature presence ("bool"), not feature frequency ("tf") [4]. However, the advanced feature frequency ("tfidf-c") is superior to the feature presence ("bool") in the proposed algorithm. Consequently, the "tfidf-c" is used in every experiment in the following subsections.

5.4 Influence of Feature Selecting

Figure **3** also displays the effectiveness of feature selecting to sentiment classification. As illustrated in Figure 3, the DF-based feature selection method can eliminate up to 50% or more of the unique substring-group features with either an improvement or no loss in classification accuracy, especially the "*tfidf-c*" curves (shown in green). In addition, Table 3 shows that all the best performances achieved by the proposed algorithm have used DF-based feature selection methods.

Based on above observations, we draw the following conclusions: the extracted substring-group features in step 1 are redundant and the feature selecting methods should be further used to eliminate the redundancy among the extracted substring-group features.

5.5 Transductive Learning vs. Inductive Learning

The following experiments show the effectiveness of using the transductive learning-based algorithm instead of inductive methods. Figure 4 gives the experimental results on the three open datasets in three different languages.

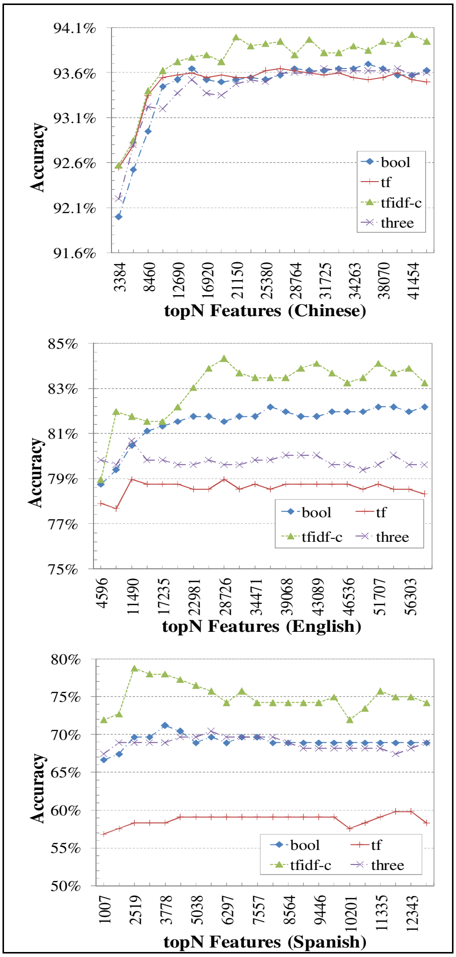

Fig. 3. The accuracies achieved by the proposed algorithm on three open sentiment datasets in different languages

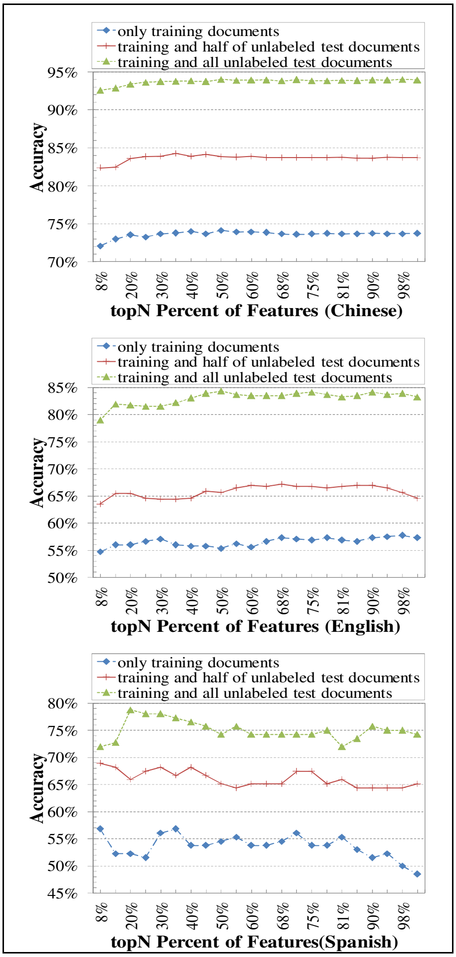

Fig. 4. Comparisons of transductive learning (green and red) and inductive learning (dark blue)

As demonstrated in Figure 4, the transductive learning (shown in green and red curves) based algorithm is well situated for sentiment classification. With the growth of the unlabeled test documents added in the suffix tree construction, the performances of transductive learning based algorithms improve significantly.

Seen from the data shown in white background in Table 3 and the dark blue curves in Figure 4, another interesting observation is that the substring-group based algorithms in inductive learning setting is inferior to the algorithms using character or word Ngrams features, which illustrate transductive learning's importance to sentiment classification from another perspective.

The reason for the improvement by transductive learning is that the more unlabeled test documents are added to the construction of the suffix tree, the more complete the structure of suffix tree becomes. This, in turn, renders the suffix tree more representative of the text corpus. This leads to extracting more representative substring-group features from the suffix tree. Result is the converted documents being more representative of the original text documents. So the unlabeled test documents' structural information used at the beginning step indirectly contributes to the feature subsumption for sentiment classification.

6 Conclusion

In this study, both feature extracting and feature selecting are incorporated into sentiment classification, and the synergetic effect of them is studied. Moreover, the proposed algorithm combines the *substring-group* features with *transductive* learning.

Experiments have been conducted on three open datasets in *three* different languages, including *Chinese*, *English* and *Spanish*. The results show that the proposed algorithm achieves better performance than the existing algorithms, without any preprocessing steps (word segmentation, stemming, etc.). Furthermore, the proposed algorithm proves to be *multilingual*, and it can be directly used for sentiment classification with any language in any encoding. In terms of term weighting approaches, the "*tfidf-c*" performs best in the proposed algorithm. Experimental results also demonstrate that the *transductive* learning based algorithm can significantly improve the classifiers' performance by incorporating the *structural information of unlabeled test documents*.

In the future, we will examine the wrong classifications to get insights on how to improve the classifier. In addition, more feature extracting methods will be explored to improve the overall performance of sentiment classification.

Acknowledgments

This work is supported by National Natural Science Foundation of China (Grant No: 60405011, 60575057 and 60875073).

References

[1] Bo, P., Lillian, L.: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2(1-2), 1–135 (2008)

[2] Bing, L.: Web data mining; Exploring hyperlinks, contents, and usage data. Springer, Heidelberg (2006)

[3] Bo, P., Lillian, L.: A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In: Proceedings of ACL (2004)

[4] Bo, P., Lillian, L., Shivakumar, V.: Thumbs up? Sentiment Classification using Machine Learning Techniques. In: Proceedings of EMNLP (2002)

[5] Ellen, R., Siddharth, P., Janyce, W.: Feature Subsumption for Opinion Analysis. In: Proceedings of EMNLP (2006)

[6] Tan, S., Zhang, J.: An empirical study of sentiment analysis for chinese documents. Expert Systems with Applications 34(4), 2622–2629 (2008)

[7] Raaijmakers, S., Kraaij, W.: A shallow approach to subjectivity classification. In: Proceedings of ICWSM (2008)

[8] Jun, L., Maosong, S.: Experimental Study on Sentiment Classification of Chinese Review using Machine Learning Techniques. In: Proceedings of IEEE NLPKE (2007)

[9] Dell, Z., Sun, L.W.: Extracting Key-Substring-Group Features for Text Classification. In: Proceedings of KDD, Philadelphia, PA (2006)

[10] Arnold, A., Nallapati, R., Cohen, W.: A comparative study of methods for transductive transfer learning. In: Proceedings of ICDM 2007 (2007)

[11] Xiaojin, Z.: Semi-supervised learning literature survey. Technical Report 1530, University of Wisconsin (2005)

[12] Sindhwani, V., Niyogi, P., Belkin, M.: Beyond the point cloud: from transductive to semi-supervised learning. In: Proceedings of ICML (2005)

[13] Joachims, T.: Transductive inference for text classification using support vector machines. In: Proceedings of ICML 1999 (1999)

[14] Vapnik, V.: Statistical Learning Theory. Wiley, NY (1998)

[15] Turney, P.D., Littman, M.L.: Unsupervised learning of semantic orientation from a hundred-billion-word corpus. Arxiv preprint cs.LG/0212012 (2002)

[16] Peter, T.: Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In: Proceedings of ACL (2002)

[17] Kim, S.-M., Eduard, H.: Determining the Sentiment of Opinions. In: Proceedings of COLING (2004)

[18] Minqing, H., Bing, L.: Mining Opinion Features in Customer Reviews. In: Proceedings of AAAI (2004)

[19] Xiaowen, D., Bing, L., Yu Philip, S.: A Holistic Lexicon-Based Approach to Opinion Mining. In: Proceedings of WSDM (2008)

[20] Alistair, K., Diana, I.: Sentiment Classification of Movie Reviews Using Contextual Valence Shifters. Computational Intelligence, Special Issue on Sentiment Analysis 22(2), 110–125 (2006)

[21] Ann, D., Khurshid, A.: Sentiment Analysis in Financial News: A Cohesion-based Approach. In: Proceedings of ACL (2007)

[22] Wan, X.: Using Bilingual Knowledge and Ensemble Techniques for Unsupervised Chinese Sentiment Analysis. In: Proceeding of EMNLP (2008)

[23] Kushal, D., Steve, L., David, P.: Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In: Proceedings of WWW (2003)

[24] Tony, M., Nigel, C.: Sentiment analysis using support vector machines with diverse information sources. In: Proceedings of EMNLP (2004)

[25] John, B., Mark, D., Fernando, P.: Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. In: Proceedings of ACL (2007)

[26] Tan, S., Wang, Y., Cheng, X.: Combining learn-based and lexicon-based techniques for sentiment detection without using labeled examples. In: Proceedings of SIGIR (2008)

[27] Ukkonen, E.: On-line construction of suffix trees. Algorithmica 14(3), 249–260 (1995)

[28] Gusfield, D.: Algorithms on strings, trees, and sequences. Cambridge University Press, New York (1997)

[29] Thorsten, J.: A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. In: Proceedings of ICML (1997)

[30] Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys (CSUR) 34(1), 1–47 (2002)

[31] Yang, Y., Pedersen, J.O.: A Comparative Study on Feature Selection in Text Categorization. In: Proceedings of ICML'97 (1997)

[32] Joachims, T.: Text Categorization with Support Vector Machines: Learning with Many Relevant Features. Springer, Heidelberg (1997)