An Information Theoretic Approach to Sentiment Polarity Classification

Yuming Lin, Jingwei Zhang, Xiaoling Wang, Aoying Zhou
Institute of Massive Computing
East China Normal University
200062 Shanghai, P.R.China
{ymlinbh, gtzhjw}@gmail.com {xlwang, ayzhou}@sei.ecnu.edu.cn

ABSTRACT

Sentiment classification is a task of classifying documents according to their overall sentiment inclination. It is very important and popular in many web applications, such as credibility analysis of news sites on the Web, recommendation system and mining online discussion. Vector space model is widely applied on modeling documents in supervised sentiment classification, in which the feature presentation (including features type and weight function) is crucial for classification accuracy. The traditional feature presentation methods of text categorization do not perform well in sentiment classification, because the expressing manners of sentiment are more subtle. We analyze the relationships of terms with sentiment labels based on information theory, and propose a method by applying information theoretic approach on sentiment classification of documents. In this paper, we adopt mutual information on quantifying the sentiment polarities of terms in a document firstly. Then the terms are weighted in vector space based on both sentiment scores and contribution to the document. We perform extensive experiments with SVM on the sets of multiple product reviews, and the experimental results show our approach is more effective than the traditional ones.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; I.2.7 [**Natural Language Processing**]: Text Analysis - *Sentiment Analysis*

General Terms

Algorithms, Experimentation

Keywords

sentiment classification, feature presentation, information theory, mutual information

1. INTRODUCTION

WebQuality '12, April 16, 2012, Lyon, France

With the prevailing of Web 2.0 techniques, more and more users prefer to share their opinions on the Web. These user-generated and sentiment-rich data are valuable to many applications like credibility analysis of news sites on the Web [7], recommendation system [6, 20], business and government intelligence [13, 1], etc. At the same time, it brings urgent need for detecting overall sentiment inclinations of documents generated by users, which can be treated as a classification problem. Sentiment analysis includes several subtasks [12] which have seen a great deal of attention in recent years: (1) Detecting if a given document is subjective or objective; (2) Identifying if a given subjective document express a positive opinion or a negative opinion; (3) Determining the sentiment strength of a given subjective document, such as strongly negative, weakly negative, neutral, weakly positive and strongly positive. In this paper we focus on the second subtask.

For identifying sentiment polarities of words, the sentiment lexicon, such as SentiWordNet [2], is an intuitive way to determine the sentiment polarities of words. The limitations of adopting this approach include: (1) The sentiment of most words or phrases are topic-dependent or domain-dependent, it is possible that one and the same word or phrase appearing in different domains can indicate different polarities. For example, "simple" is negative in book domain while being positive in electronics domain. (2) Many words always have multiple meanings, it is difficult to determine which one should be chosen in current text without context. Thus, more sophisticated approaches are required for evaluating the sentiment polarities of words.

Besides the methods based on sentiment lexicons, an influential work is done by Pang et al. [17], which compared the the performances of several machine learning methods with different feature presentations and reported that the SVM with unigram presence worked best. We treat it as a baseline since we also pay attention to multiple feature presentations. In the bag of words framework, the documents are always converted into vectors based on predefined features including feature types (unigram, bigram, etc.) and feature weighting method (frequency, presence, TF*IDF and its variants, etc.), which is critical for classification accuracy. But most of these methods did not take into account the term's sentiment orientation when its weight in a document is evaluated. If this factor is associated with term's score in documents, the classification accuracy should be improved, which is demonstrated in our experiments.

To integrate a term's sentiment polarity into its weight,

Martineau [10] proposed Delta TF*IDF (ΔTF*IDF for brevity), in which the number of positive (negative) documents with term t is used to evaluate the term's sentiment orientation. For ΔTF*IDF, the weight $V_{t,d}$ of term t in document d is evaluated as follow:

$$V_{t,d} = C_{t,d} * log_2(N_t/P_t) \quad (1)$$

where $C_{t,d}$ is the number of times term t occurs in document d, N_t is the number of documents with negative label in the training set with term t, and P_t is the number of documents with positive label in the training set with term t.

We argue it is not sufficient since the distribution of the term in documents in the training set are ignored when a term's relationship with the positive (negative) label is quantified. For example, there are three documents in total, one of them is positive and the others negative. Term t occurs k times ($k \geq 2$) in the positive one, and once in only one of the negative documents. Intuitively, term t is more likely to express positive sentiment. But its score of ΔTF*IDF is always zero, and then the contribution of term t will be ignored. Therefore, the weighting methods are required to be more sophisticated in such cases. In this paper, we present an improved feature weighting method, in which the mutual information is applied on quantifying the term's sentiment polarity, the term's frequency in a document describes its contribution to this document.

This paper makes the following contributions:

1. Applying information theory for the sentiment polarity classification problem;
2. Proposing an improved feature weighting method, which quantifies the terms sentiment scores by mutual information;
3. Performing a series of experiments on several kinds of feature presentation methods for sentiment classification in multiple product reviews. The results show the proposed approach is more effective than the traditional ones.

The rest of this paper is organized as follows. Section 2 provides an overview of the related work. Our solution is presented in detail in Section 3. A series of experiments are conducted to demonstrate the effectiveness of our solution in Section 4. Finally, we conclude our work in Section 5.

2. RELATED WORK

Sentiment classification of documents can be treated as a special text categorization. Many feature weighting methods were proposed for topic-based text categorization , such as TF*IDF, information gain [19] and term strength [23], etc. But more effective methods are required in sentiment classification because sentiment expressions are more subtle. We outline the related work on sentiment classification based on identifying the term sentiment polarity or not.

Firstly, most researches on sentiment classification have focused on training machine learning algorithms to classify reviews. Pang et al [17] compared eight feature presentations methods including unigram, bigram, the combination of unigram and bigram with feature frequency or presence in review polarity classification implemented by Naive Bayes, maximum and SVM respectively. And the SVM with information of feature presence based on unigram outperformed the others in their experiments, whose accuracy achieved 82.9%. Two years later they used minimum cut to filter the object content from their reviews, and then they trained and tested SVM classifier on their trimmed reviews [16]. On the basis of unigram and bigram, [11] expanded the features with frequent word sub-sequences and dependency sub-tree, which improved the classification performance. The above work did not focus on the term sentiment polarity, which started from a statistical perspective entirely.

Secondly, the usage of sentiment lexicons, such as SentiWordNet[2], HowNet[4] and WordNet[12], is a straight way to identifying the sentiment polarity of word [18][22]. The sentiment scores of words are pre-evaluated which always range from -1 to 1. To construct such a lexicon, some propagation strategies have been applied in existing work. For example, [5] used a set of seed adjectives with clear sentiment polarity to grow this set by searching their synonym/antonym in WordNet. More recently, Lu et al tried to construct a context-dependent sentiment lexicon by linear programming [9]. But the sentiment polarity of a document is more than just a linear sum of the sentiment scores of the terms occurring in that document [14]. Thus, we weight the terms with their sentiment score determined in SentiWordNet together with frequency, and treat this weighting method as another baseline in our experiments. Turney [21] suggested determining the polarity of a word or phrase by measuring its point-wise mutual information with some seeds like "excellent" and "poor" in unsupervised sentiment classification, which encountered the same problem as using sentiment lexicon essentially on one hand. One the other hand, the effectiveness of this way depends on an external corpus which is used to measure the relation between the terms in text and the seeds. Such requirement limits the adaptability of this way in some sense.

However, some researches identified the sentiment polarity of document based on terms polarities without the external resources. Paltoglou and Thelwall [15] compared multiple variants of classic TF*IDF schema adapted to sentiment analysis, and emphasized that expressing sample vectors with emotional information via supervised methods is helpful for predicting sentiment polarity. Moreover, Matineau and Finin[10] proposed ΔTF*IDF to present document data for sentiment polarity classification, by which the importance of discriminative terms can be identified and boosted. [24] made a comparative study on five feature selection, including term selection based on document frequency, information gain, mutual information, χ^2 test and term strength, they reported that the χ^2 test was the most effective which was evaluated by the equation (2):

$$\chi^2(t,l) = \frac{N \times (A_l D_l - C_l B_l)^2}{(A_l + C_l) \times (B_l + D_l) \times (A_l + B_l) \times (C_l + D_l)} \quad (2)$$

where A_l is the number of times t and l co-occur, B_l is the number of times the t occurs without l, C_l is the number of times l occurs without t, D_l is the number of times neither l nor t occurs, and N is the total number of documents. In two-category (such as l_1, l_2) setting, the value $\chi^2(t,l_1)$ of term t and label l_1 is equal to the $\chi^2(t,l_2)$, because $A_{l_1} = B_{l_2}$, $B_{l_1} = A_{l_2}$, $C_{l_1} = D_{l_2}$ and $D_{l_1} = C_{l_2}$ always hold. Thus, the χ^2 test doesn't work in sentiment polarity classification.

Consequently, a more sophisticated method is acquired for measuring the sentiment polarity of terms. We enhance

the relevance of terms and polarity labels with mutual information in this paper, which is verified to be effective for increasing the classification accuracy in our experiments.

3. THE PROPOSED APPROACH

Our approach consists of two parts: first, we capture the sentiment tendency of the terms via evaluating their mutual information with polarity labels; second, the contribution of terms to a document is determined, which will be combined with the former to weight the terms in this document.

3.1 Identifying Term's Sentiment Tendency

Given a training set $S = \{\langle s_1, l_{s_1}\rangle, \ldots, \langle s_n, l_{s_n}\rangle\}$, where s_i denotes the i^{th} training sample and l_{s_i} the polarity label of s_i, and a set of testing samples $U=\{u_1,\ldots,u_r\}$ without labels. Our task is to predict the polarity labels of all testing samples. In this subsection, we present how to apply mutual information to capture a term's sentiment polarity.

In probability theory and information theory, the mutual information can capture the difference between the joint distribution on (X, Y) and the marginal distributions on X and Y. Moreover, it is a quantity that measures the mutual dependence of two random variables. Formally, the mutual information of two discrete values is evaluated as follow:

$$MI(x, y) = \log_2 \frac{p(x, y)}{p(x)p(y)} \tag{3}$$

where $p(x, y)$ is the joint probability of x and y, $p(x)$ and $p(y)$ are the marginal probability of x and that of y respectively.

Now we quantify a term's relationship with each label by mutual information. Given N labeled samples, A is the number of times term t and label l co-occur, B is the number of times term t occurs without label l, C is the number of samples with label l but not include term t. Thus the mutual information $MI(t, l)$ of t and l can be evaluated by:

$$MI(t, l) = \log_2 \frac{p(t, l)}{p(t)p(l)} \approx \log_2 \frac{A \times N}{(A + B) \times (A + C)} \tag{4}$$

In this paper, we focus on two types of labels, the positive label l_p and the negative label l_n. For the example discussed in Section 1, i.e., term t occurs k times ($k \geq 2$) in a positive document and once in only one of two negative documents. In this case, $MI(t, l_p) = log_2(3/(k + 1))$ and $MI(t, l_n) = log_2(3/(2(k + 1)))$. Term t is more likely to express positive sentiment, since $MI(t, l_p) > MI(t, l_n)$ holds.

In general, if term t is positive, the value of $MI(t, l_p)$ is relatively high and $MI(t, l_n)$ is relatively low according to the formula (4). Thus, The sentiment score $S_{MI}(t, l_p)$ of term t on l_p can be derived from a linear combination of $MI(t, l_p)$ and $MI(t, l_n)$, the $S_{MI}(t, l_n)$ is similar to $S_{MI}(t, l_p)$:

$$\begin{cases} S_{MI}(t, l_p) = \alpha MI(t, l_p) + (1 - \alpha)(-MI(t, l_n)) \\ S_{MI}(t, l_n) = \alpha MI(t, l_n) + (1 - \alpha)(-MI(t, l_p)) \end{cases} \tag{5}$$

where $0 \leq \alpha \leq 1$ and it is a weighting parameter which reflects the contributions of $MI(t, l_p)$ and $MI(t, l_n)$.

To measure the goodness of term t in a global feature presentation, we integrate $S_{MI}(t, l_p)$ and $S_{MI}(t, l_n)$ into one sentimtne score $S_{MI}(t)$ with formula (6):

$$S_{MI}(t) = \begin{cases} |S_{MI}(t, l_p)| & \text{if } S_{MI}(t, l_p) > S_{MI}(t, l_n) \\ 0 & \text{if } S_{MI}(t, l_p) = S_{MI}(t, l_n) \\ -|S_{MI}(t, l_n)| & \text{if } S_{MI}(t, l_p) < S_{MI}(t, l_n) \end{cases} \tag{6}$$

Table 1: The top 15 positive unigrams and negative ones in reviews on kitchen appliances and books ($\alpha = 0.7$)

kitchen appliances		books	
positive	**negative**	**positive**	**negative**
awesome	not_waste	nationalism	disappointing
loves	defective	Trotsky	weak
amazing	refund	masterpiece	flat
quiet	probe	awesome	predictable
favorite	returned	favorites	poorly
ease	junk	investigation	repetitive
impressed	worst	imagination	not_waste
Lodge	not_unit	explores	stupid
collection	dangerous	rare	barely
remember	failed	Crichton	useless
sizes	not_recommend	vivid	unrealistic
durable	followed	illustrated	boring
cleans	ring	copies	endless
excellent	told	Vietnam	zero
Dutch	candy	sons	pathetic

If $S_{MI}(t, l_p) > S_{MI}(t, l_n)$ holds, term t tends to express positive sentiment in the current domain, the overall sentiment score of t, $S_{MI}(t)$, is set to a positive value. On the contrary, if term t tends to negative sentiment, its $S_{MI}(t)$ value should be negative. For example, considering the term t with $MI(t, l_p) = 5$ and $MI(t, l_n) = 0.2$; it means term t is more inclined to express positive sentiment, then $S_{mi}(t) = 5$. Notably, in the case of $MI(t, l_p) = 0.5$ and $MI(t, l_n) = -10$, term t is more relative to positive label rather than the negative one according to formula (4), thus $S_{mi}(t) = 0.5$. But if $MI(t, l_p)$ is equal to $MI(t, l_n)$, it means term t has no contribution for identifying the polarities of documents.

Thinking about another extreme case: term t appears k times ($k \geq 2$) in one positive document without occurring in the rest positive documents. At the same time, term t appears once in each negative document. For this case, term t should tend to negative sentiment. In our solution, the value of $S_{MI}(t, l_p)$ is smaller than that of $S_{MI}(t, l_n)$, which means term t is more relevant to negative label. Thus, our approach can capture the correct sentiment inclinations of terms in a corpus.

To make an intuitive understanding, Table 1 shows the top 15 positive unigrams and negative ones for reviews on kitchen appliances and books in our real-world dataset respectively. We can observe that most of these terms reflect correct sentiment inclinations intuitively. Here, the term "not_waste" denotes a tag "not_" is appended to the term "waste", which will be described in subsection 4.2. When someone does not like something, he or she can often say "Don't waste your time on ...". Thus, we should properly process the negatory words. Some outliers like "Lodge" and "Trotsky" etc. will be discussed further in subsection 4.3.

3.2 Identifying document's sentiment Polarity

In this subsection, we present how to apply the sentiment scores of terms to identify the sentiment polarities of documents automatically.

To evaluate the contribution of term t to document d, we combine the frequency of t in d and the sentiment score $S_{MI}(t)$. Thus the weight $V(t, d)$ for term t in document d is

defined as:

$$V(t,d) = tf_{t,d} \times S_{MI}(t) \quad (7)$$

where the $tf_{t,d}$ is the frequency of term t in document d.

Some researches like [17] reported the presence features is comparable and even better than the frequency feature due to the sparsity of opinionated words. But our key observation in pre-experiment is that it depends on the analyzed domains, and even it achieves the worst performance sometime. Thus we apply the frequency of terms on capturing their contribution to the document.

The procedure of sentiment polarity classification using mutual information is described in algorithm 1. Firstly, for each term t occurring in samples of the training set, its sentiment score S_{MI} value is evaluated with formula (4), (5) and (6). Secondly, the $V(t, s_i)$ value of each term t in each training sample s_i is determined with formula (7). Let symbol $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ denote the training set using the proposed approach, where $\mathbf{x}_i$ is the vector of the i^{th} sample. And the vector $\mathbf{x}_i$'s component $x_{i_{t,s_i}}$ refers to the $V(t, s_i)$ of term t and training sample s_i. $Y = (\mathbf{y}_1, \ldots, \mathbf{y}_r)$ is similar to X, excepting $\mathbf{y}_j$ denotes the vector of testing sample. Then, a classifier C is generated by training on the set X, and C is used to generate the set $L = \{l_1, \ldots, l_r\}$ of sentiment polarity labels of testing samples.

Algorithm1 Sentiment Polarity Classification based on *MI*
Input: the training set $S = \{\langle s_1, l_{s_1}\rangle, \ldots, \langle s_n, l_{s_n}\rangle\}$,
the testing set $U = \{u_1, \ldots, u_r\}$
Output: predicted labels set $L = \{l_1, \ldots, l_r\}$
1: $L = \varnothing$, $S^{'} = \varnothing$;
2: for each term t occurring in S do
3: evaluate $S_{MI}(t, l_p)$ and $S_{MI}(t, l_n)$ according to formula (4) and (5);
4: determine $S_{MI}(t)$ with formula (6);
5: for i=1 to n do
6: generate vector $\mathbf{x}_i$ for s_i according to formula (7);
7: $S^{'} = S^{'} \cup \{\langle \mathbf{x}_i, l_{s_i}\rangle\}$;
8: train the classifier C on $S^{'}$;
9: for i=1 to r do
10: generate vector $\mathbf{y}_i$ for u_i according to formula (7);
11: $l_i = C(\mathbf{y}_i)$;
12: $L = L \cup \{l_i\}$;
13: return L;

4. EXPERIMENTS

4.1 Experimental Setting

To validate the effectiveness and robustness of the proposed approach, a real-world dataset[1] reorganized by Blitzer et al. [3] is prepared for our experiments, which consists of reviews of books (**B** for short), DVDs (**D**), electronics (**E**) and kitchen appliances (**K**) from Amazon[2]. The reviews marked with 4 or 5 stars are labeled with a positive label, and those with 1 or 2 stars are labeled with a negative label. Each product domain contains 1000 positive and 1000 negative reviews. Five-fold cross validation is applied in our experiments. All tests used LIBSVM[3] with a linear kernel function, the rest parameters remained the default values.

[1]http://www.cs.jhu.edu/∼mdredze/datasets/sentiment/
[2]http://www.amazom.com
[3]http://www.csie.ntu.edu.tw/∼cjlin/libsvm/

Table 2: A 2×2 Contingence Table

Doc Number	actual positive	actual negative
predict positive	a	b
predict negative	c	d

We focus on two types of features used commonly in sentiment classification: unigram and bigram. For each one, we compare our approach with other weighting methods including frequency, presence and ΔTF*IDF respectively:

1. **tf*senti**: SentiWordNet 3.0 is used to determine the sentiment score of a term, the term is weighted by the product of its frequency and its sentiment score.
2. **frequency**: The term frequency is regarded as its weight.
3. **presence**: Considering whether the term occurs in a document.
4. **delta tfidf**: Described in Section 1 (ΔTF*IDF).
5. **tf*MI** (our): Weighting terms with formula (7).

The accuracy is applied on evaluating the effectiveness of proposed approach in our experiments, which is computed as follow:

$$accuracy = \frac{a+d}{a+b+c+d} \quad (8)$$

where a, b, c and d are described in Table 2.

4.2 Preprocessing

The data set is prepared according to the following steps: (1) We remove all punctuations but retain the stop words. (2) All unigrams with length less than 3 are omitted. No stemming or lemmatizing is used because they are detrimental to classification accuracy[8]. (3) Like the work done in [17], we remove the negatory words from reviews and append the tag "not_" to the words following the negatory word in a sentence. For instance, the sentence "It doesn't work smoothly." would be altered to become "It not_work not_smoothly.".

4.3 Results and Discussion

In this subsection, we present the experimental results to demonstrate the effectiveness of our approach.

The first experiment concerns the effect of the sentiment score S_{MI} presented in subsection 3.1. Recalling the results in Table 1, almost all of the top 15 sentiment scores of positive unigrams and negative ones in reviews of books and kitchen appliances reflect strong sentiment direction correctly. Now we focus on the analysis of some outliers. The "Lodge" in column 1 of Table 1 always express the neutral sentiment without context. But the "Lodge" is a famous manufacture, whose products on kitchen appliances including Dutch ovens retrain a lot of popular credibility. There are 10 reviews total on the products of Lodge in our dataset, and 90 percent of them are positive. That is why the term "Lodge" builds stronger relationship with positive label. Seven of the ten reviews on sharpeners are labeled as positive, and the left three negative. Thus, the term "sharpeners" is more closely related to the positive label. The terms "probe" and "candy" are similar to this case, the former is

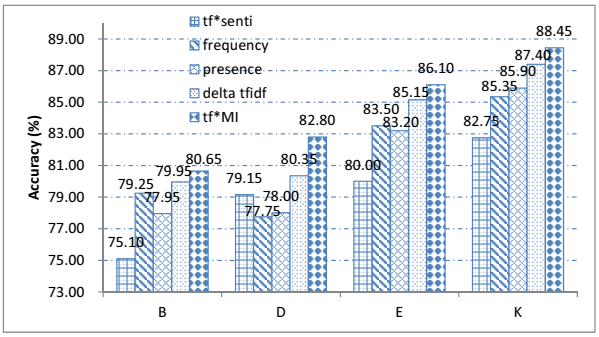

Figure 1: Accuracy comparisons based on unigram

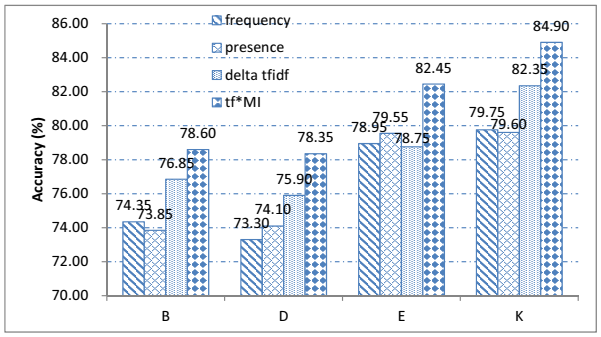

Figure 2: Accuracy comparisons based on bigram

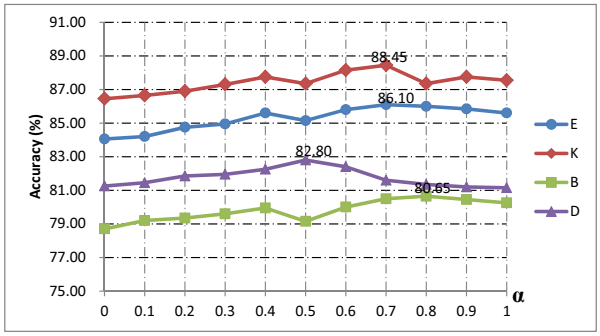

Figure 3: Accuracies of tf*MI based on unigram under varying value of α

included in 15 negative reviews and 1 positive review, and the later occurs in 6 negative reviews and in 1 positive review. In our dataset the Dutch ovens received wide praise, but "ovens" does not occur in table 1 due to "oven" used by reviewers sometimes and no stemming applied in our experiments. We omit such discussion on DVDs and electronics domains, which are similar to the cases discussed above.

From the results in Table 1 for book domain, "Trotsky" and "Crichton" are two writers whose books gained recognition in our dataset. The most reviews on books referring to the term "Vietnam" are positive (five positive reviews and only one negative review), the term "sons" is similar to it. For the term "zero" in books, the negative reviews contain some negative information including "zero information", "zero interesting anecdotes", while "zero" is often regarded as the neutral one considered in isolation. Thus, our approach captures term's sentiment tendency correctly.

The unigram is one of the most commonly used feature type in text classification. The second experiment concerns the effectiveness of **tf*MI** based on unigrams. Figure 1 shows the results of accuracy comparisons on four product domains. The value of α on unigram for each domain is shown in Figure 3. Our approach, **tf*MI**, achieves the best performance in all product domains. On the one hand, both **delta tfidf** and **tf*MI** outperform the other three weighting methods invariably. On the other hand, **tf*MI** is better than **delta tfidf** in all domains, especially the accuracies of **tf*MI** is about 2.5% higher than that of **delta tfidf** in **D**. Thus, both **tf*MI** and **delta tfidf** are effective ways for unigram to sentiment classification and the second half, $log_2(N_t/P_t)$, of formula (1) can also capture the sentiment polarities of terms effectively, but our approach is more effective. At the same time, **tf*senti** performs poorly in total, because a word often express several meanings. When it is considered solely, it is difficult to determine its sentiment polarity in a document. We observe that the feature presence is not always superior to the feature frequency, that is true in **D** and **K** but false in **B** and **E**. Comparatively, it seems the feature frequency is better than feature presence based on the average accuracy in all domains.

Figure 2 shows the comparisons on the weighting methods based on bigram. The value of α on bigram for each domain is shown in Figure 4. Like the previous experiment on unigram, our approach, **tf*MI**, achieves the best performance in all domains, and the accuracy is improved more significantly comparing whit the results in Figure 1. It is worth noting that **delta tfidf** is worse than the frequency and presence in electronics domain for the bigram, the **delta tfidf** is not very stable in some sense. Comparatively, our approach always achieves the best performance in all domains.

Further, we observe that the classification accuracies on reviews of electronics and kitchen appliances are higher than those in books and DVDs domains clearly. The sentiment expressions on books or DVDs are often more subtle than those of general products. For instance, a review of a book is written as following:"When I read this book, I can't conceal my rage on the leading man, his ugly personality make me sick.". The terms "rage", "ugly" and "sick" always express intensive negative emotion, while we concern them without context. But the reviewer is praising this book indeed.

At last we concern the impact of varying the parameter α based on unigram and bigram respectively. Recalling the

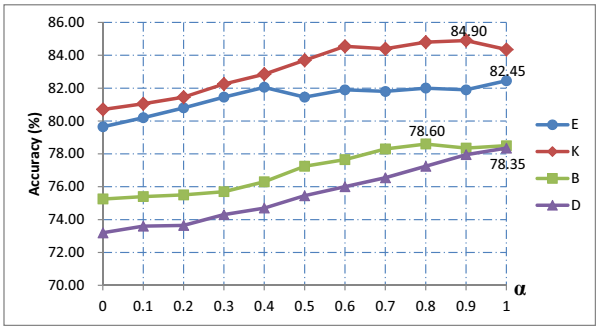

Figure 4: Accuracies of tf*MI based on bigram under varying value of α

discussion in Section 3, term t's score of polarity label l is linear combination of two parts: t's mutual information with l and that with the opposite of l. As shown in Figure 3 and Figure 4, when **tf*MI** achieves the best accuracy, the value of α always locates at [0.5, 1]. Thus, the relation between t and l seems to be more importance than that between t and the opposite of l, that is consistent with our intuitive. For unigram, the value of α should range from 0.5 to 0.8, and from 0.8 to 1 for bigram. Notably, whatever is the value of α set, accuracies of our approach is higher than that of **delta tfidf** in almost all cases as long as $\alpha \geq 0.5$ holds.

5. CONCLUSION

Sentiment classification has seen a great deal of attention in recent years, on which the bag of words of framework is widely applied. In such settings the predefined feature type and weighting method are crucial to classification accuracy. In this paper we introduce information theory into the sentiment polarity classification, and propose an improved feature weighting method for sentiment polarity classification of documents, in which sentiment polarities of terms are identified correctly, which are expressed as sentiment scores evaluated based on mutual information. To measure term's contribution to a document, its frequency in this document is integrated into our solution. In a series of experiments, our approach achieves the best performance in a real-world dataset including multiple product reviews comparing with the traditional weighting methods.

6. ACKNOWLEDGMENTS

This work was supported by the 973 project (No. 2010CB3 28106), NSFC grant (No. 61033007 and 60925008), Program for New Century Excellent Talents in University (No. NCET-10-0388) and Shanghai Leading Academic Discipline Project grant (No. B412).

7. REFERENCES

[1] A. Abbasi. Affect intensity analysis of dark web forums. In *Proc. of Intelligence and Security Informatics*, pages 282–288, 2007.

[2] S. Baccianella, A. Esuli, and F. Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *the Seventh conference on International Language Resources and Evaluation*, pages 2200–2204, May 2010.

[3] J. Blitzer, M. Dredze, and F. Pereira. Biographies, bollywood, boo-boxes and blenders: Domain adaptation for sentiment classification. In *Proc. 45th ACL*, pages 440–447, June 2007.

[4] Z. Dong and Q. Dong. Hownet - a hybrid language and knowledge resource. In *Natural Language Processing and Knowledge Engineering,2003*, pages 820–824, Oct. 2003.

[5] M. Hu and B. Liu. Mining and summarizing customer reviews. In *KDD'04*, pages 168–177, August 2004.

[6] J.Tatemura. Virtual reviewers for collaborative exploration of movie reviews. In *Proc. of Intelligent User Interfaces*, pages 272–275, January 2000.

[7] Y. Kawai, Y. Fujita, T. Kumamoto, J. Jianwei, and K. Tanaka. Using a sentiment map for visualizing credibility of news sites on the web. In *Proceedings of the 2nd ACM workshop on Information credibility on the web*, pages 53–58, November 2008.

[8] E. Leopold and J. Kindermann. Text categorization with support vector machines. how to represent texts in input space? *Machine Learning*, 46(1):423–444, January 2002.

[9] Y. Lu, M. Castellanos, U. Dayal, and C. Zhai. Automatic construction of a context-aware sentiment lexicon: An optimization approach. In *WWW 2011*, pages 347–356, March 2011.

[10] J. Matineau and T. Finin. Delta tfidf:an improved feature space for sentiment analysis. In *Proceedings of the Third International ICWSM Conference*, pages 258–261, May 2009.

[11] S. Matsumoto, H. Takamra, and M. Okumura. Sentiment classification using word sub-sequences and dependency sub-trees. In *PAKDD 2005*, pages 301–311, May 2005.

[12] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. Introduction to wordnet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–312, 1990.

[13] G. Mishne and N. Glance. Predicting movie from blogger sentiment. In *AAAI Symposium on Computational Approaches to Analysing Weblogs*, pages 155–158, 2006.

[14] K. Moilanen and S. Pulman. Sentiment composition. In *RANLP-07*, pages 378–382, September 2007.

[15] G. Paltoglou and M. Thelwall. A study of information retrieval weigheing schemes for sentiment analysis. In *Proc. 48th ACL*, pages 1386–1395, July 2010.

[16] B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimun cuts. In *Proc. 42th ACL*, pages 271–278, July 2004.

[17] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using maching learning technique. In *Proc. 7th EMNLP*, pages 79–86, July 2002.

[18] A.-M. Popescu and M. Pennacchiotti. Detecting controversial event from twitter. In *CIKM'10*, pages 1873–1876, October 2010.

[19] J. Quinlan. Induction of decision tree. *machine learning*, 1(1):81–106, 1986.

[20] L. Terveen, W.Hill, B.Amento, D.McDonald, and J. Creter. Phoaks: A system for sharing recommendations. *Communications of the Association for Computing Machinery*, 40:59–62, 1997.

[21] P. D. Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proc. 40th ACL*, pages 417–424, July 2002.

[22] Y. Xia, L. Wang, K.-F. Wong, and M. Xu. Sentiment vector space model for lyric-based song sentiment classification. In *Proc. ACL-08:HLT*, pages 133–136, June 2008.

[23] Y. Yang. Noise reduction in a statistical approach to text categorization. In *SIGIR'95*, pages 256–263, July 1995.

[24] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *Proc. 14th ICML*, pages 412–420, July 1997.