Multiple Factors-based Opinion Retrieval and Coarse-to-Fine Sentiment Classification

Shu Zhang, Wenjie Jia, Yingju Xia, Yao Meng, Hao Yu
Fujitsu Research & Development Center, China
zhangshu@cn.fujitsu.com, wj_jia@cn.fujitsu.com, yjxia@cn.fujitsu.com
mengyao@cn.fujitsu.com, yu@cn.fujitsu.com

Abstract

With more and more reviews on the web, browsing through a mass of the related reviews becomes a heavy work. How to effectively analyzing and organizing these reviews attracts more attention. Opinion mining deals with the computational treatment of opinion, sentiment, and subjectivity in documents has received a lot of attention in natural language processing and information retrieval research community. This paper focuses on retrieving the opinion documents and giving their sentiment orientation. The topic relevant opinion documents are measured with a sentiment model, combining the existing knowledge and statistic information. Opinion lexicon is fused in the model to measure the sentiment strength. Multi-level sentiment analysis approach is proposed to find the topic related sentiment information. Supervised method and unsupervised method are adopted to solve the document level and phrase level classification. Our experimental results on COAE show the effectiveness of the proposed methods, and prove the feasibility of classifying orientation at varying levels of granularity by fused different methods.

Keywords

Opinion retrieval; sentiment classification; opinion mining.

1 Introduction

With large amounts of user-generated information resource, how to find the interesting information, analyze and extract useful information has received a lot of attention in natural language processing and information retrieval research community. The opinion-related applications include product or movie review analysis, political opinion polls, and advertisement analysis, etc.

A series of topic symposiums and evaluation sessions on opinion mining have appeared in TREC and NTCIR (Ounis et al, 2008; Seki et al, 2008). There is a growing interest in finding out opinions from the web data. In the TREC Blog Track, one task is to retrieve relevant blog posts in topic and identifying documents with opinionated contents. A similar mission in NTCIR multilingual opinion analysis task is to extract opinion related properties from the given documents for each topic.

Opinion retrieval is one task of searching and finding opinions over general topics with the aim of presenting documents containing personal opinions towards the given query. It is to rank the document according to the balance of subjectivity and relevance. How to measure the sentiment and how to combine opinion with relevance are key problems in research.

Contrast with fact-based documents classification, sentiment classification often has relatively few classes, which include binary classification or ordinal categories(Pang et al, 2008). Opinion-oriented information extraction more focuses on the opinion expression (e.g., holder, type, strength) than topic, which means it could be generalized well across different domains. These differences make opinion-oriented tasks appear easier than fact-based textual analysis. However, this is far from the truth. Fact-based documents differ considerably between two topical different documents due to the topic related word. Sentiment might often be expressed in a subtle manner, leading it difficult to be identified by any of the isolated sentence or document's terms. Sentiment is quite context sensitive, the same expression could indicate different sentiment in different domains. Sequential information and discourse structure also seem more important. All these might be the challenges and charms for sentiment analysis.

In this paper, we focus on problems of opinion retrieval and sentiment classification: first search and rank documents over the topic (query) and sentiment strength, then classify their sentiment on the given topic. We start from the general information retrieval to get the topic related documents, then construct the opinion model to measure sentiment strength, and utilize a typical linear combination of relevance score and opinion (sentiment) score to re-rank these documents. In this stage, we focus on how to measure the sentiment strength, how to merge the existing knowledge and statistic information into the sentiment model. For sentiment classification, we try to analyze the sentiment at varying levels of granularity in order to capture the real topic related opinion and improve the performance.

The remainder of the paper is organized as follows: section 2 describes the related work. Section 3 gives our approach on opinion retrieval. Section 4 presents the methods of SVMs-based and phrase-pair extraction based sentiment classification. Section 5 gives the experiments and results. Finally section 6 summarizes this paper.

2 Related Work

Opinion mining is first proposed by Dave et al (2003) to process a set of search results for a given item, generate a lists of product attributes and aggregate opinions about each of them. Sentiment analysis parallels with opinion mining, appeared in the papers on classifying reviews and analysis of evaluative text(Turney et al, 2002; Pang et al, 2002). They deal with the computational treatment of opinion, sentiment, and subjectivity in documents.

Opinion retrieval is a task to find relevant and opinionate documents according to a user's query. It generally adopts a two-stage process, topic relevance based search and sentiment based re-ranking. The topic relevance is important, many examples show the existing methods on document opinion ranking provide no improvement over mere topic relevance ranking(Ounis et al, 2006). This also indicates the current opinion ranking methods has more extent to be improved. At present, one major method to find subjective content is based on sentiment lexicon and weights them by calculating term frequency(Mishne et al, 2006; Oard et al, 2006). There are also some works on modeling the topic and sentiment in a unified way and measuring the term weight for sentiment analysis. A generation model that unifies topic relevance and opinion generation by a quadratic combination is proposed by Mei et al (2007), which induces the relevance-based ranking as the weighting factor of the lexicon-based sentiment ranking function, generally flexible in practice. Emphasis on how to represent and evaluate the weight of sentiment terms, a sentiment analysis model is proposed and proved effectiveness by Kim et al(2009).

Here we also adopt the typical linear combination of relevance weight and sentiment weight, pay more attention to the sentiment model with our existing knowledge and statistic information.

Sentiment classification is to determine whether sentiment expressed is positive or negative in the given granularity unit. Sentiment classification could be focused on different granularity, from document level, sentence level, to phrase level, for different applications with different needs. Classifying the sentiment expressed in the document is a canonical machine learning task. With large annotated corpus labeled of sentiment orientation, supervised methods such as SVMs and Na ïve Bayes perform well on this task(Pang et al, 2002). It considers that the problem of classifying documents not by topic, but by overall sentiment, determining whether a review is positive or negative. SVMs has been testified to do the best in the performance. A representative unsupervised algorithm (Turney et al, 2002) is to classify reviews by averaging semantic orientation of phrases in the review. The semantic orientation is calculated by comparing the mutual information between the given phrase and the word “excellent” or “poor”. Li et al (2009) also try the semi-supervised models aiming to utilize unlabeled corpus.

Here we consider the orientation of the document towards the given topic. There is a problem that some documents represent author’s attitudes over multiple issues rather than a single issue. The author might first give positive attitude on one issue and then express negative sentiment on the other issue. So the traditional sentiment classification on document level focusing on overall positive or overall negative is not suitable for this task. We classify the sentiment at varying levels of granularity-- document level and phrase level, and utilize supervised method and unsupervised method to confirm the document’s orientation towards the given topic.

3 Opinion Retrieval

The task is to find relevant and opinionate documents to a query, how to measure the topic related information and sentiment information are two key issues in opinion retrieval. We consider these from the topic relevance analysis and sentiment analysis aspects separately. Topic relevance analysis is more studied in IR field, so we focus on the sentiment aspect in this task.

In order to rank the document by their relevance, we utilize the topic relevance retrieval method, find and rank documents according to its extent of topic relatedness. Then we re-rank the documents combined with the opinion scores.

The final scores of the documents are calculated by a linear combination of relevance score and opinion score, which is shown below.

$$S(d,q) = \lambda \cdot S_{topic}(d,q) + (1-\lambda) \cdot S_{opinion}(d,q)$$

Where $S_{topic}(d,q)$ is to estimate topic relevance, $S_{opinion}(d,q)$ is to estimate sentiment strength, λ is the linear combination weight, measures their proportion in equation, set as 0.3 empirically.

3.1 Topic Relevance Analysis

Topic relevance analysis is the first step to get what you want to find. Traditional methods have proved effectiveness in searching and ranking the document by given query. So IR model is adopted to find and weight the topic relevant documents. Language model is one of the famous ones in the field of IR. With query q as input, retrieved documents are ranked based on the probability that the document's language model would generate the terms of the query.

$$\begin{aligned} S_{topic}(d,q) &= \sum\nolimits_{w\in q} \ln P_{Dir}(w \mid d) \\ &= \sum\nolimits_{w\in q} \ln \frac{tf(w,d)+\mu P_{ML}(w \mid C)}{|d|+\mu} \end{aligned}$$

Where $tf(w,d)$ is the frequency of word w in the document *d*, $P_{ML}(w \mid C)$ is probability of word w in the document collection C，|*d*| is the length of document *d*, μ is the constant.

3.2 Sentiment Analysis

Sentiment analysis is one of the key issues in the task of opinion retrieval. It aims to measure the sentiment strength of the relevant documents retrieved by IR models.

There is a hypothesis: one document contains more opinion words, and these words are more associated with the given topic (query), then the document is considered to have more sentiment strength. So we consider it from query, content words and sentiment words, and define opinion score as the probability of document d contains relevant sentiment op to the given query q. It is shown in the following.

$$\begin{aligned} S_{opinion}(d,q) &\equiv p(d \mid op, q) \\ &\propto p(op, q \mid d) \\ &= \sum_{w\in d} p(op, q \mid w)\cdot p(w \mid d) \\ &= \sum_{w\in d} p(op \mid w)\cdot p(q \mid w)\cdot p(w \mid d) \end{aligned}$$

Here, *op* is a variant, defined as 1, if it has sentiment, otherwise 0. Assuming uniform prior probabilities of document *d*, query *q* and *op*, and conditional independence between *q* and *op*, the opinion scores change to $p(op, q \mid d)$. For simplicity, we assume the documents represented as a bag-of-word and all the words are uniformly distributed.

The final opinion score is combined with three factors: the probability of word w to be a sentimental word, the likelihood of query q given the word w, and probability of word w being generated by document *d*.

$p(op \mid w)$ is to measure the probability of word to be a sentimental word, which is calculated based on the existing sentiment lexicon.

General opinion words lexicon, which contains about 5,000 comment and sentiment words. They are oriented to all the fields. Each opinion words have a sentiment orientation and extent with it.

For example, <安全(safety) +2>. Here, +, - and 0 are used to represent positive, negative and neutral orientation.

The extent is ranged from 0 to 2 with the increasing sentiment degree.

$$p(op \mid w) = \begin{cases} |EvaluationValue|/3 \\ 0 \end{cases}$$

Where $|EvaluationValue| \in [0, 2]$, it is the strength of word w given in the lexicon. We assume the strength of the words indicate the probability of word w to be a sentimental word.

For $p(q \mid w)$, we utilize the web statistic information to capture the association of the word w with the given query *q*. The more two words co-occur with each other, the more they have a relation in semantic.

We simplify to estimate the value as the following:

$$p(q \mid w) = \frac{N(q \cap w)}{N(q)}$$

$N(q)$ is the count of retrieved documents with the query word q, $N(q \cap w)$ is the count of retrieved documents with both the query q and word w .

The last part $p(w \mid d)$ is to evaluate importance of a word in a document. It could be classified to the traditional IR problem. So we also measure it with the language model as shown in section 3.1.

4 Sentiment Classification

A closer look at our sentiment classification task, it aims to classify the orientation of document on the given topic. Previous work on sentiment classification is judged from the overall positive or negative in one document. Our task is to judge sentiment of the document associated with the query.

For documents focused on one topic, their orientation could be judged from the overall positive or negative sentiment. However, the overall sentiment are not the exactly attitude to the given topic when documents include more than one issue, especially multiple issues with different attitudes.

So we identify the sentiment from coarse to fine grain. Figure 1 shows the architecture of our method. We get the overall sentiment from the document level. And we also analyze the sentiment towards the topic from the phrase level. For the sentiment analysis, we utilize the opinion lexicon to catch the opinion words and its orientation in documents. At last, we combine these two results. Here, we adopt the supervised method for the document level analysis, and unsupervised method for the phrase level analysis. Feature set includes both general features and sentimental features, which are selected by training.

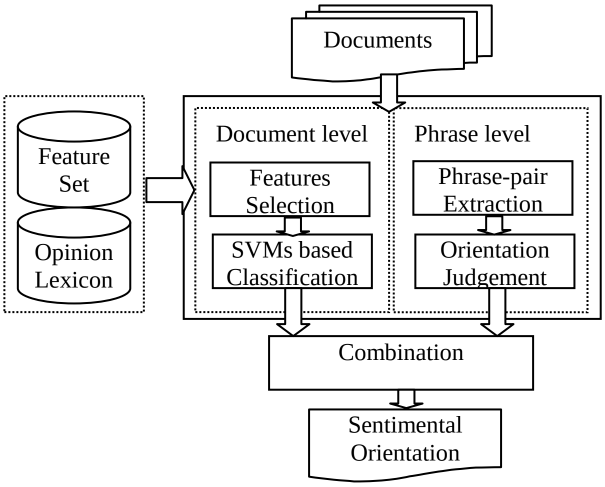

Figure 1. Sentimental classification architecture

4.1 Support Vector Machines Based Method

Support Vector Machines which are based on the Structural Risk Minimization principle from computational learning theory, have been shown be highly effective at traditional text categorization. In the two-category case, the basic idea is to find a hyper-plane based on the training data. There are many hyper-planes that might classify the data. The best hyper-plane is the one that represents the largest separation, or margin, between the two classes, which means the distance from it to the nearest data point on each side is maximized. If such a hyper-plane exists, it is known as the maximum-margin hyper-plane. Figure 2 shows the main idea.

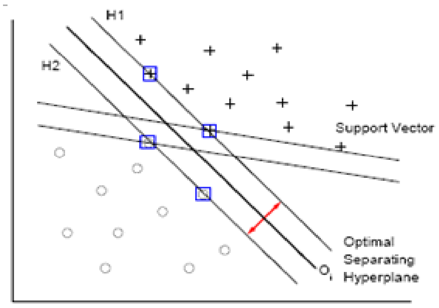

Figure 2. Support vector machines

Here we adopt SVMs to classify the orientation on document level, orientation includes positive, negative and neutral. Documents are represented by the vector space model, and tf-idf is chosen as the weight of the features. The training corpus contains about 2,000 positive, 2,000 negative, and 1,000 neutral documents. Features selected from the corpus include two types, general words and sentimental words. We choose about 6,000 general words and 2,000 sentimental words as features.

Feature selection is important for classification performance and has received more attention from the researchers. There are some methods considering the importance of features to the document from the different aspects, such as document frequency, CHI statistics, mutual information et al.

We adopt the category analysis based features selection method (Wang et al, 2006), which selects features optimally according to the feature contribution to each class. The main thought of the method is that a good feature subset is the one that contains features highly correlated with the class, which could be measured by the document frequency and term within document frequency in a class, the larger document frequency and term frequency values in a class, the stronger relation between the feature and the class. The advantages are using both word and document frequency to measure the feature importance, and inducing variance mechanism to mine the latent category information.

General features and sentimental features are selected by this features selection method separately. General features are the words associated with topic content aspect. Based on general opinion word lexicon, the sentimental words are firstly selected and sentimental features are chosen. Then eliminating sentimental words, general features are chosen. General features and sentimental features are utilized to measure the document from topic relatedness, sentimental orientation and sentimental strength.

4.2 Phrase-pair Extraction Based Method

Besides document level analysis, we go deeply to finer grain--phrase level to classify the sentiment orientation.

As observed, people often express their sentiment around the topic in sentences. The opinion words around the topic are more likely having the real sentiment expression. So we construct the topic word and its associated sentimental word as phrase-pair, and want to capture the document's sentiment from them.

At present, the cross sentence analysis is difficult, the performance is low. So we simplify to find the phrase-pair in the sentences, which contain both the topic words and sentimental words. For the association between phrase and opinion words, we adopt the nearest vicinity match method to confirm the associated sentimental word.

We assign the orientation of the phrase-pair from multi-level: sentence, context and sentimental word level. Orientation judgement is defined in the following way.

$$O(Pair_i) = IsO(S) \cdot IsN(a_i) \cdot Sign(o_i)$$

Here, *IsO(S)* is a two value function, which judges the orientation by sentence level. If the sentence is judged to have no sentiment, then its value is 0, else is 1. At present, we only consider hypothesis sentence. We think this type of sentence doesn't express sentimental information.

IsN(a_i) is also a two value function, which considers the sentiment transition by emotional adverbs or phrase. It is on the context level to consider the orientation. If there is such a word around the topic or sentimental word in a region, the value is -1, else is 1.

Sign(o_i) is directly the orientation of associated sentimental word. Its value is defined as -1 for negative, 0 for neutral and 1for positive orientation.

Both the information about a_i and o_i are determined by looking at the lexicon, which has been described in section 3.2.

The sentiment of the document towards the topic on phrase level is calculated by the following equation. It is measured by the ratio between the counts of the positive phrase-pair and negative ones. The orientation is judged by the pre-defined threshold, if the sentiment value is greater than the threshold, it is positive, otherwise negative.

$$Sentiment(d) = \frac{N(pos)}{N(neg)}$$

The calculating of document's sentiment could also adopt other measurement, such as mean value or machine learning methods.

4.3 Combination

In order to obtain a better classification results, the results gotten from two proposed methods are combined. We choose the SVMs based method as the baseline, introduce the phrase level analysis to modify the result. It has been implemented as follows:

- Classify the orientation of the documents by SVMs-based method, assign the document as positive, negative and neutral.
- The document assigned with neutral orientation is the final result, without any more analysis
- The document assigned with positive and negative orientation, is modified its result according to orientation judged by phrase-pair extraction based method.

With the experience, phrase-pair extraction based method performs better on the documents with positive and negative orientation than neutral orientation. It is more effective to capture the topic and opinion words in one sentence than in document.

Opinion words found in a document may be associated with other topic or things, do not express any sentiment to the given topic. However, they are all considered as factors in document level analysis, this caused the performance lower. Phrase-based judegment reduce the scope to the sentence, so it may grasp the sentiment towards the topic more accurately.

However more often the topic and opinion words are not appeared in the same sentences, phrase-pair based judgement gets higher precision and lower recall compared with SVMs-based method. So we combine document level and phrase level analysis to get the real sentimental orientation to the given topic.

5 Experiments

The data used in experiments are provided by COAE (The second Chinese Opinion Analysis Evaluation), which was held in 2009, aimed to enable researchers to participate in large-scale experiments and evaluations, make each researcher’s result comparable and promote the related techniques in Chinese opinion analysis.

There are five tasks in COAE 2009, including sentiment analysis and opinion mining at word level, sentence level, and document level.

We participated in task 5, which is to retrieve and rank the topic relevant opinion documents, and classify their orientation on the given topic.

The corpus consists of 40,000 documents and 50 topics. The topics are related to person name, event, product, etc. Our systems submitted in COAE are introduced in Table 1. For the limit of submitted runs, we mainly compare the performance of SVMs-based and the combined methods on sentiment classification. The performance of the opinion retrieval is testified by comparing with other submitted systems.

System ID	Methods
Run1	SVMs
Run2	SVMs + phrase-pair

Table 1. Corresponding table of submitted systems

Table 2 presents the evaluation results on opinion retrieval. We also give the Avg. and Max. values in the task. There are 7 participants, 14 runs in the task. Our systems are named as Run1 and Run2. Table 2 shows the results considers about the first 100 and 300 ranked documents.

System ID	100		300	
	MAP_rel	MAP_senti	MAP_rel	MAP_senti
Run1(2)	**0.6009**	**0.4186**	**0.6298**	**0.4272**
Median	0.5130	0.3453	0.53695	0.3523
Best	0.6009	0.4186	0.6298	0.4272

Table 2. Opinion retrieval results

In Table 2, MAP_rel measures the ability of searching and ranking document merely based on the topic relevance without sentiment. MAP_senti measures the ability not only on topic relevance but also on sentiment strength. We get the best performance on both the topic relevance retrieval and opinion retrieval, which proves the system architecture is reasonable and the proposed methods are effective.

In Table 2, it shows that the proposed methods are performed consistently in both the first 100 and 300 documents retrieval.

In order to have a whole view of the performance of the all submitted systems. Figure 3 and Figure 4 show the 14 system results on Map_rel and MAP_senti measurement. Here, the results are gotten by considering the first 300 documents. System ID 1 and 2 are our systems. In Figure 3, most systems have the similar performances on topic relevance retrieval, participants adopt the traditional IR methods, the values of MAP_rel are ranged from 0.5 to 0.6298. Considering the sentiment, the performances of all systems decrease at some extent, which are shown in Figure 4. How to measure the sentiment strength of documents is the focus of this task. The comparison between the values MAP-rel and MAP_senti gives a good guidance to find the direction to improve the sentiment measurement effectively.

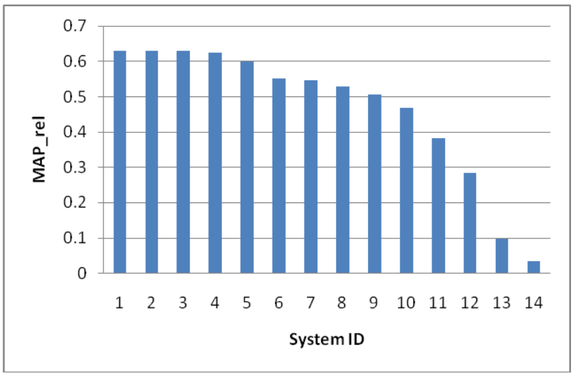

Figure 3. MAP_rel results of 14 system

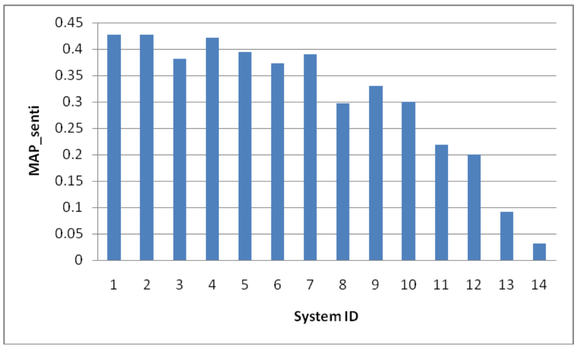

Figure 4. MAP_senti results of 14 system

For the performance of sentiment classification, it is testified based on the retrieval result, shown in Table 3 and 4. That means the right results are the topic related opinion documents with correct orientation. Here R-accuracy considers the first 100 and 300 ranked documents. Table 3 and Table 4 show that the method combined SVM and phrase-pairs extraction gets the best performance in all submitted systems. That proves the proposed method is feasible and effective.

RunID	R-accuracy	Precision	Recall	F1
Run1	0.173766	0.0725	0.342	0.105021
Run2	**0.221337**	0.1053	**0.4**	0.141275
Avg	0.1761	0.1239	0.338	0.155236
Max	0.221337	0.1859	0.396	0.219588

Table 3. Sentiment classification evaluation results (100)

RunID	R-accuracy	Precision	Recall	F1
Run1	0.173766	0.0724828	0.341713	0.105021
Run2	**0.221337**	0.105291	**0.396062**	0.141275
Avg	0.1761	0.1239125	0.338402	0.1552355
Max	0.221337	0.185867	0.396062	0.219588

Table 4. Sentiment classification evaluation results (300)

Compared Run1 with Run2, the values of combined method are all higher than those of SVMs-based method, these show that fine grain analysis is a feasible way to improve the performance on the document level sentiment classification. With the revising by the phrase-pair based orientation judgement, the orientation is more accurate. At the same time, the recall of Run2 is also higher than that of Run 1. That also shows that fusion of multiple classifiers approach for opinion analysis is a good way to improve the performance.

Look closer to the testing corpus, which has a certain amount of documents having multi-topics. This is one reason that the induced fine grain analysis achieves improvement in the performance.

The precision of the method is lower than the Avg. value, which indicates the proposed method need more improvement, such as the parameter setting, the phrase-pair extraction and the combination strategy.

Though the proposed methods get good results in COAE task, the performance is still low both on the opinion retrieval and sentiment classification. This is only primary work on opinion retrieval, and need more future work to improve the performance.

6 Conclusion

In this paper, we probe into the problems of opinion retrieval and sentiment classification with the given topic. In opinion retrieval, we measure the sentiment strength combined the existing knowledge with statistic information, and merge them into one model. In sentiment classification, we analyze from document to phrase level, get the orientation of the documents towards the given topic. Our experimental results on COAE are encouraging, which prove the effectiveness of the proposed techniques, the feasibility of classifying orientation at varying levels of granularity. Phrased-pair based orientation judgement shows effective to improve the results gotten from the document level analysis. This work is primary, more room needs to improve in the further work.

7 References

Ounis I., Macdonald C. and Soboroff I., Overview of the TREC-2008 Blog Track. In: Proc. 17thText REtrieval Conference, 2008, pp. 15-27.

Seki Y., Evans D. K., et al., Overview of Multilingual Opinion Analysis Task at NTCIR-7. In: Proc. NII Test Collection for IR Systems, 2008, pp.185-203.

Pang B. and Lee L., Opinion Mining and Sentiment Analysis. Foundation and Trends in Information Retrieval, vol.2, 2008, pp. 1-135.

Dave K., Lawrence S. and Pennock D. K., Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. In: Proc. 12th International World Wide Web Conference, 2003, pp. 519-528.

Turney P., Thumbs up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In: Proc. 40th Association for Computational Linguistics, 2002, pp. 417-424.

Pang B., Lee L. and Vaithyanathan S., Thumbs Up? Sentiment Classification Using Machine Learning Techniques. In: Proc. Empirical Methods in Natural Language Process, 2002, pp.79-86.

Ounis I., Rijke M. de, Macdonald C., et al., Overview of the TREC 2006 Blog Track. In: Proc. 15thText REtrieval Conference, 2006, pp.17-31.

Mishne G., Multiple Ranking Strategies for Opinion Retrieval in Blogs. In: Proc. 15thText REtrieval Conference. Maryland(2006). http://trec.nist.gov/

Oard D., Elsayed T., Wang J. and Wu Y., TREC-2006 at Maryland: Blog, Enterprise, Legal and QA Tracks. In: Proc. 15thText REtrieval Conference. Maryland (2006) http://trec.nist.gov/

Mei Q. Z., Ling X., Wondra M., Su H. and Zhai C. X., Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs. In: Proc. 16th International World Wide Web Conference, 2007, pp.171-180.

Kim J., Li J. J.and Lee J. H., Discovering the Discriminative Views: Measuring Term Weights for Sentiment Analysis. In: Proc. 47th Annual Meeting of the ACL and 4th IJCNLP of the AFNLP, 2009, pp. 253-261.

Li T., Zhang Y.and Sindhwani V., A Non-negative Matrix Tri-factorization Approach to Sentiment Classification with Lexical Prior Knowledge. In: Proc. 47th Annual Meeting of the ACL and 4th IJCNLP of the AFNLP, 2009, pp. 244-252.

Wang Q., Guan Y., Wang X. L.and Xu Z. M., A Novel Feature Selection Method Based on Category Information Analysis for Class Prejudging in Text Classfication. In: International Journal of Computer Science and Network Security, vol. 6, No.1A, pp. 113-119, 2006.