Accessing Positive and Negative Online Opinions*

Hanhoon Kang, Seong Joon Yoo, and Dongil Han

School of Computer Engineering, Sejong University, 98, Gunja, Gwangjin Seoul, Korea
sjyoo@sejong.ac.kr, sjyoo@sejong.ac.kr, dihan@sejong.ac.kr

Abstract. Nowadays, an increasing number of people review the comments on each item before they will purchase the commodities and services offered by online shopping malls, Internet blogs, or cafés. However, it is somewhat challenging to routinely read trough all of the comments. The purpose of this study is to introduce some methods to classify the positive or negative review pertaining to the blog comments on a movie written in Korean. For this purpose, a variety of algorithms was used to classify the reviews and allow feature-selection by applying the traditional machine learning method for classifying literature.

Keywords: opinion mining, machine learning, text categorization.

1 Introduction

With the recent flurry of activity by people who participate in Internet blogs, cafés, and other communities, Internet users are readily uploading the UCC (User Created Contents) onto their personal blogs, and frequently found in the UCC is the review written by the user on his or her own purchase or the services they used. Also, for the online shopping mall, it is available to present a comment on the product purchased but increasingly the often compulsory limitation to a concise input of comments should deteriorate the reliability of those reviews. Therefore, the process of selecting a purchase may primarily involve searching through the comments of other users on the product being considered. However searching and reading through the scores of comments can be tedious labor. In addition, choosing to look only at the reviews on the 1st page can lead to a misguided purchase due to the potential prejudice pertaining to the single opinion suggested by a single user. In this regard, this study will draw out the method to classify the positive/negative reviews in a specific comment, as the initial stage of the course to summarize and show the reviews presented by a number of users, using a supervised machine learning algorithm. Many studies[1][2][3][4], have been conducted on this subject, most of which targeted the documents from an English-speaking culture using a professional data set of review documents. In this study, reviews on movies, commodities, and newspaper articles were used as input data applying the extension of WordNet to the words, ‘positive and negative’. Data,

[*] This research was conducted with the support of ‘Seoul R&BD Program(10581)’.

C. Stephanidis (Ed.): Universal Access in HCI, Part III, HCII 2009, LNCS 5616, pp. 359–368, 2009.
© Springer-Verlag Berlin Heidelberg 2009

that was specific to the comments on movies in a number of Korean blogs, was collected using a blog crawler. The study results on the blog review documents were generated by Naïve Bayesian, Score Function, and Naïve Bayes methods of Support Vector Machine, WEKA Library, and Information Gain, OddRatio, Mutual Information, TFIDF, LogTFIDF and Cosine(IS), Klosgen as the Feature-selection method.

2 Related Works

Classifying Positive/Negative Online Opinions

[1] is about a study that applied a traditional document classification method adopting the machine learning technique. In this study, 700 positive and 700 negative review documents on movies were collected from the IMDb (Internet Movie Database) to be put under the Support Vector Machine, Naïve Bayes algorithms. Feature words were selected through various methods, including unigrams and bigrams, and the weight of those words was calculated using the term frequency and presence methods. The results showed the highest performance of 82.9% in the SVM classification algorithm when the weight of the words were estimated via the presence method.

[2] is about defining the Score Function classification algorithm. Experiment [3] was conducted with the review documents for 7 types of commodity items with the number of review documents reaching up to more than 1000. The words were selected though extraction, in the form of bigrams or trigrams, with the result showing performance of 84.6%-88.3%, derived from a serious of classification algorithms including naïve Bayesian and SVM.

[4] is about conducting the classification on 4 different topics based on emotional word-phrases. In this study, 170 positive reviews and 140 negative reviews were collected from Epinions[6]. The parts of speech were extracted from the documents collected, which are consistent with particular patterns, with the PMI (Pointwise Mutual Information) that was modified before it was applied. In the modified PMI formula, using the ‘near’ operator in the Alta Vista search engine led to determination of several phrases of semantic orientation (SO) that fit certain patterns included in a particular document. Addition of SO for several phrases equals to the semantic orientation of a single document. The result indicated more than 80% of classification accuracy for those topics of ‘Automobiles’ and ‘Banks’.

[7] was an experiment to classify the positive and negative reviews of the users’ comments on tourist sites using the supervised machine learning method. User reviews on 7 tourist destinations introduced by the portal website yahoo.com were collected with classification algorithms including SVM, Naïve Bayes, and a dynamic language model classifier. Information Gain was used as the weight algorithm, which was also used as feature-selection in SVM and Naïve Bayes. The results were represented according to the number of training data, showing the highest classification performance in SVM when the positive and negative data set indicated 300 each.

[5] is an experiment for the classification of emotions (positive and negative) with data sets that were collected from Korean commodity reviews, movie reviews, and

newspaper articles. For the emotional characteristics, the WordNet from English-speaking cultures was expanded and applied. SVM was used as the classification algorithm, with TF-IDF and TF-ISF as the feature-selection method.

In classifying the online plane documents, the previous studies were usually focused on the documents found in English-speaking cultures, using the contents of some professional review sites. The few studies on the emotional classification of Korean documents also used the traditional document classification methods. The same methods were adopted in this study.

In spite of the fact that in [5], the SVM algorithm and the TF-IDF and TF-ISF feature-selection methods were applied, following the extraction of positive/negative features with the emotional words through the English version of WordNet, this study is focused on extracting emotional words including adjectives, adverbs, or verbs from the blog comments about movies. The major difference between Korean and English in extracting those emotional words is the morphologic aspect. In this study, it was necessary to determine which part of speech the words belong to using a morpheme analyzer.

Also, various types of classification algorithms and feature-selection methods were applied to the results in an attempt to figure out whether the application of a certain type of algorithm and feature-selection method leads to a balanced result of both positive and negative aspects. Also, using the blog crawler, the documents collected were divided into 10 data sets for each of which 5-fold cross validation was applied to achieve relatively precise results for the experiment.

Text Categorization

In [8], the Gini index was adopted as a newly experimented algorithm to select the feature values. In this algorithm, the traditional Gini index is modified for use in the classification of documents. The experiment results were put under comparison analysis through the feature-selection methods including Information Gain, Cross Entroy, and CHI Square, in addition to the Gini index, Support Vector Machine, and k-Nearest Neighbor.

In [9], the experiment was conducted through a new type of feature-selection algorithm, multi-criteria ranking method, and Reuters-21578 data set because of the known importance of selecting the feature values.

[10] redefines the term commonly known as Mutual Information using the Pointwise Mutual Information method, while with this same algorithm the feature values were extracted and processed into lab results.

In [11], various types of feature selection method were utilized for the document classification lab through kNN and LLSF algorithms.

[12] introduces LOGTFIDF, the improved version of LOGTFIDF as an algorithm for feature selection method in the text classification. The results are presented for the feature values selected through either TFIDF or LOGTFIDF. The data set used here is Reuters-21578.

In [13], Naïve Bayes, Rochio, C4.5, k-NN and SVM methods were used for the classification of documents. Especially the use of polynomial and rbf in the kernel function of SVM with the parameter verified values which resulted in excellent performance compared with other algorithms.

3 Machine Learning Algorithms for Classifying Positive/Negative Online Opinions

Support Vector Machine

The optimal borderline that SVM[15] is trying to find is the bordering side considering the maximum Margin. The spots of data supporting the bordering side (one blue-filled circle and 2 red squares in Figure 2) are called Support Vector.

The reasons SVM classifier considers the maximum margin when finding the bordering side for classification are: First, the availability of stable operation with the best performance; second, the maximum margin leading to the minimization of small errors found around the classification borders. The most central of the classification border sides considering the maximum margin is called the Optimal Separating Hyperplane, which is represented in formula (1), where, when f(x)>0, the circle class is +1, while f(x)<0 resulting in the circle class(-1). This formula can be put under repetitive executions to generate a training model.

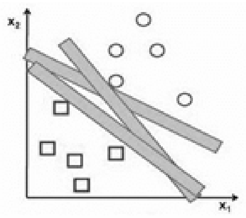


Fig. 1. Search of Hyperplane

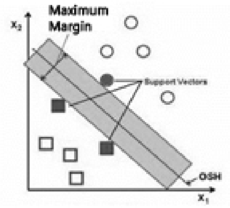


Fig. 2. Decision of Hyperplane

$$f(x) = \sin(w^T x + b) \tag{1}$$

In formula (1), w and b are the parameters of Hyper-plane with f(x) = ± 1. However, since the Input Vectors are not always linearly separable, SVM will use the mapping function (Φ) to escalate the vector space of lower level to an upper level, and using the Support Vector, conduct the classification by identifying the maximum Margin. Here, the different kinds of Kernels are applied.

$$K(u,v) = \Phi(u) \bullet \Phi(v) \tag{2}$$

In formula (2), the Kernel functions include Gaussian RBF, Sigmoid, Polynomial, and Inverse multi-quadric Kernel. And formula (3) as a discretion function, predicts the class for a given vector.

$$f(x) = \sin(\sum_{i=1}^{N} \hat{a}_i y_i x_i^T x + \hat{b}) \tag{3}$$

Naïve Bayes Classifier

Naïve Bayes Classifier [16]is applied to the document classification based on the Bayes Probability. The formula in which the document D is allocated to C is as follows when the word lists are comprised of W(w_1-w_d) for the newly document, i.e., the unclassified document D.

$$C = P(c_j)\prod_{i=1}^{d} P(W_i \mid c_j)$$

P(c_j) is a pre-probability that is classified in category c_j, and P(w_i |c $_i$) is a post-probability in which w_i is included in c $_I$. In the Naïve Bayes Classification, it is assumed that the event in which the words appear in the document is independent.

When the group of documents to be examined is small, P(w_i |c $_i$) can be inaccurate. If there is no word included in the learning category of the document data to be classified, the probability value will be 0. The following formula can be used to calculate the post-probability and the cases in which 0 is included in the value of P (w_i|c $_i$).

$$P(w_i \mid c_j) = \frac{n_{ij} + 1}{n_j + k_j}$$

N_j is the total number of words within c_j and N_{ij} is the frequency of appearance of the word w_i. k_j is the number of words in the classification c_j. When the calculation is made with this value, it is possible to acquire a value similar to that in the previous calculation of the Bayesian Probability while the value resulting in 0 can be prevented [21].

WEKA Naïve Bayesian

The WEKA [17] Library is an open source library that was implemented by Java as a machine learning algorithm package. Here, a variety of algorithms are provided, and this paper has adopted the Naïve Bayesian algorithm. The reason the Naïve Bayesian algorithm of WEKA was adopted in this study is because the feature values like SVM are supposed to be input in the format of a particular file, namely the ARFF [21], in the normal Naïve Bayes and WEKA, Naïve Bayes and WEKA Naïve Bayesian are to be compared with each other.

Score Function Classifier

The Score Function Classifier [20] is a kind of supervised learning method. This is a review document, in which the individual words are extracted with their scores obtained through the following probability formula,

$$score(t_i) = \frac{\Pr(t_i \mid C) - \Pr(t_i \mid C^{'})}{\Pr(t_i \mid C) + \Pr(t_i \mid C^{'})}$$

where the total scores of the individual words appearing in the document d_i is $eval(d_i) > 0$, it is classified as the positive review, otherwise the negative review.

$$eval(d_i) = \sum_j score(t_j)$$

$\Pr(t_i \mid C)$ that is used to calculate $score(t_i)$ is a conditional probability where t_i is produced on condition that class C is generated; and here, C is the positive review.

On the contrary, $\Pr(t_i \mid C')$ is a conditional probability where t_i is produced on condition that class $C^{'}$ is generated; and here, C is categorized into the negative review. In this study, the probability values are calculated with the following formulas.

$$\Pr(t_i \mid C) = \{ \text{(Frequency of } t_i \text{ in C+} \mu \text{)} / |C| \}$$

$$\Pr(t_i \mid C') = \{ \text{(Frequency of } t_i \text{ in C}' + \mu \text{)} / |C'| \}$$

|C|=The sum of the frequency of the individual words in the n number of positive reviews.
|C'|=The sum of the frequency of the individual words in the n number of negative reviews.

The reason why μ is added to the frequency of t_i is because when the frequency equals 0, a negative calculation error occurs, therefore, the value 0.01, which is the modified value for the Laplace estimator, is applied.

4 Feature Selection for Classifying Online Opinions

In this study, the documents were comprised with the weight vectors as in the traditional classification methods. Rather than nouns, the only words used in this study were emotional words, which are categorized into adverbs, verbs, and adjectives. The total number of emotional words extracted was 5,620. Based on the different emotional words, the feature-selection method was applied to the Support Vector Machine and WEKA Library Naïve Bayesian algorithms; and probability methods to Score Function and Naïve Bayesian algorithms. The methods used in this study include Information Gain (IG) [11], Mutual Information (MI) [11], TFIDF [19], LogTFIDF [12], Cosine (IS)[18], and Klosgen[18]. And IG, MI, and TFIDF methods are the feature-selection methods frequently adopted in the classification studies. Cosine(IS) and Klosgen used in this study are normally adopted for the estimation of objective measures of rules in Association Rule Mining [18].

The decision to adopt these methods is ascribed to the fact that the intention of this study was to utilize the various features selection methods. The experimental results show similar outcomes for both positive and negative categories. Since [12] shows the result of LogTFIDF known to improve the linearity problems which function unfavorably in TF, also in this study it has produced similar outcomes for both categories in WEKA-LogTFIDF.

5 Experiments

Dataset

In this paper, a blog crawler was implemented to collect Korean blog (http://section.blog.naver.com) documents from which documents containing the comments on movies were extracted. Since the blog comments had a consistent structure, the extraction was favorable. For example, the comments include not only general information, such as the director and cast, but also a ranking system (1-5 stars). In this report, the rankings of 1 and 2 were designated as negative, and 4 and 5, positive. As a result, the positive reviews totaled 16,800 and the negative reviews reached 1680, equivalent to 1/10 of the total positive cases. To balance the ratio, the positive cases were set to the same 1680. However, only 10 data sets have been included in this paper. With 1680 positive reviews, each data set of negative reviews was arranged and coupled with each positive review.

Estimation of Performance and Measures

As described earlier, in this study, 10 data sets were arranged, for each of which 5-fold cross validation was conducted. The mean values of these were calculated by the following estimation measures [20].

$$\text{Precision (P)} = \text{TP} / (\text{TP} + \text{FP})$$

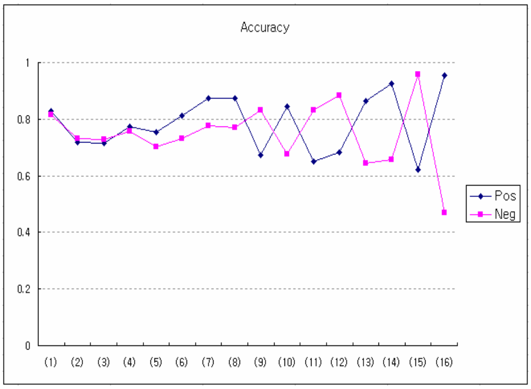
$$\text{Recall (R)} = \text{TP} / (\text{TP} + \text{FN})$$

$$F_1\text{-Measure} = (2*R*P) / (R+P)$$

Table 1. Meanings of Acronyms

Prediction Classification	TRUE	FALSE
TRUE	TP	FP
FALSE	FN	TN

Results



(1) SVM-OddRatio, (2)WEKA-LogTFIDF, (3) WEKA-Klosgen
(4) WEKA-IG, (5) WEKA-TFIDF, (6) WEKA-MI, (7)SVM-MI
(8)SVM-IG, (9)SVM-Klosgen, (10) WEKA-OddRatio
(11) Naïve Bayes, (12) SVM-Cosine(IS), (13)WEKA-Cosine(IS)
(14)SVM-TFIDF, (15)Score Function, (16) SVM-LogTFIDF

As seen in the results, when SVM was applied to the positive review along with LogTFIDF as the feature-selection method, the performance was the highest; in the negative review, the application of Score Function resulted in the best performance. However, in terms of SVM-LogTFIDF or Score Function, the results of classification for the opposite review were not so valid. In this study a similar classification result between both positive and negative reviews was assumed to indicate a more reliable classification method. The similar performances in classification for both categories

Table 2. Accuracy

Method	Pos Avg	Pos Std.Dev.	Neg Avg	Neg Std.Dev
(1)	0.828	0.061	0.816	0.017
(2)	0.720	0.057	0.733	0.023
(3)	0.716	0.033	0.729	0.054
(4)	0.773	0.045	0.757	0.046
(5)	0.753	0.033	0.702	0.055
(6)	0.812	0.018	0.733	0.067
(7)	0.875	0.010	0.777	0.068
(8)	0.875	0.048	0.771	0.028
(9)	0.672	0.042	0.832	0.027
(10)	0.844	0.001	0.675	0.001
(11)	0.652	0.031	0.833	0.048
(12)	0.682	0.027	0.885	0.046
(13)	0.865	0.011	0.644	0.059
(14)	0.927	0.021	0.658	0.061
(15)	0.622	0.035	0.958	0.053
(16)	0.953	0.028	0.470	0.041

were found in SVM-OddRatio, WEKA-LogTFIDF, WEKA-Klosgen, WEKA-IG, etc. Considering the balance between classification performances for both positive and negative reviews, the SVM-OddRatio indicated a higher performance.

Table 2 shows the precision as well as standard deviation values. This study tried to look at the stability of the classification task using the distance from the mean values to each of the observed values since 5-fold cross validation on the 10 data sets was conducted.

The following results are those for F-Measure reflecting the Recall and Precision. In this result, the F-Measures appeared to be similar among SVM-OddRatio, WEKA-LogTFIDF, WEKA-Klosgen, WEKA-IG, WEKA-TFIDF, WEKA-MI, SVM-MI, and SVM-IG, however, SVM-MI and SVM-IG showed the highest F-Measure values, when the balance between the classification performances of both phases of reviews was taken into consideration.

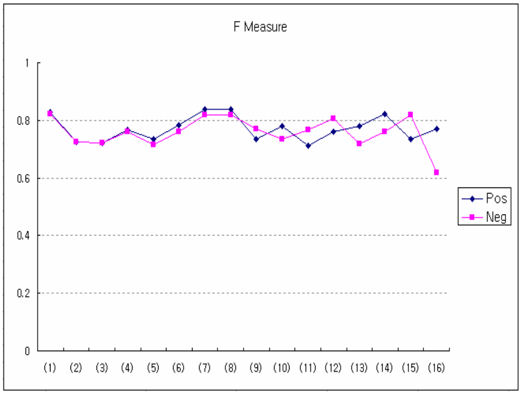


Table 3. F-Measure

Method	Pos Avg	Pos Std.Dev.	Neg Avg	Neg Std.Dev
(1)	0.828	0.037	0.822	0.037
(2)	0.725	0.022	0.725	0.038
(3)	0.721	0.030	0.722	0.035
(4)	0.767	0.001	0.762	0.001
(5)	0.735	0.021	0.716	0.040
(6)	0.782	0.028	0.760	0.036
(7)	0.839	0.032	0.818	0.041
(8)	0.839	0.032	0.818	0.041
(9)	0.733	0.038	0.771	0.027
(10)	0.780	0.029	0.734	0.041
(11)	0.711	0.045	0.766	0.027
(12)	0.761	0.043	0.804	0.030
(13)	0.780	0.031	0.718	0.046
(14)	0.822	0.025	0.759	0.051
(15)	0.734	0.049	0.818	0.025
(16)	0.771	0.024	0.617	0.068

6 Conclusion

In the Web 2.0 environment, the users have actively participated in Internet cafes, blogs, and other communities for communication. However, these information can be utilized in various fields. The comments attached to each product in which users, describing what they have felt through the whole process of purchasing, may be usefully explored in the marketing area. It is not easy for the users to read every single comment on a certain purchase, so summarizing the points of those comments will facilitate the purchase process as well as help the enterprises to utilize them in the marketing of their products by comparing them with other makers. In this regard, this study implemented the crawler function on the Korean blog documents to collect relevant data while applying various feature-selection methods and classification algorithms for comparison analyses. As a result of the study, some algorithms were proven to be effective specific to either positive or negative opinions, while some algorithms performed well for both items. It may be necessary, in future studies, to make an effort to explore the Natural Language Processing for the Korean language for more accurate classification.

References

1. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment Classification using Machine Learning Techniques. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 79–86 (2002)
2. Dave, K., Lawrence, S., Pennock, D.: Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. In: Proc. of the 12th Intl. World Wide Web Conference (WWW 2003), pp. 512–528 (2003)
3. Liu, B.: Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. Springer, Heidelberg

4. Turney, P.D.: Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 417–424 (2002)
5. Hwang, J., Ko, Y.: A Korean Sentence and Document Sentiment Classification System Using Sentiment Features. Journal of Korean Institute of Information Scientists and Engineers (KIISE): Computing Practices and Letters 14(3), 336–340 (2008)
6. `http://www.epinions.com/`
7. Ye, Q., Zhang, Z., Law, R.: Sentiment classification of online reviews to travel destination by supervised machine learning approaches. Expert Systems with Applications, 1–9 (2008)
8. Shang, W., Houkuan, Zhu, H., Lin, Y., Qu, Y., Wang, Z.: A novel feature selection algorithm for text categorization. Expert System with Application 33, 1–5 (2007)
9. Doan, S., Horiguchi, S.: An Efficient Feature Selection Using Multi-Criteria in Text Categorization. In: Proceedings of the 4th International Conference on Hybrid Intelligent Sysmstems, HIS 2004 (2004)
10. Xu, Y., Jones, G., Li, J., Wang, B., Sun, C.: A Study on Mutual Information-based Feature Selection for Text Categorization. Journal of Computational Information Systems 1(2), 203–213 (2005)
11. Yang, Y., Pedersen, J.O.: A Comparative Study on Feature Selection in Text Categorization. Proceedings of 14th International Conference on Machine Learning, 412–420 (1997)
12. Liao, C., Alpha, S., Dixon, P.: Feature Preparation in Text Categorization. Aritificial Intelligence White Papers, Oracle
13. Joachims, T.: Text Categorization with Support Vector machine: Learning with Many Relevant Features. In: Proceedings of 10th European Conference on Machine Learning, pp. 137–142 (1998)
14. Kang, H., Yoo, S.J.: SVM and Collaborative Filtering-Based Prediction of User Preference for Digital Fashion Recommedation Systems. IEICE Transaction on Information and Systems E90-D(12), 2100–2103 (2007)
15. Hsu, C.-W., Chang, C.-C., Lin, C.-J.: A practical guide to support vector classification
16. Isa, D., Lee, L.H., Kallimani, V.P., RajKumar, R.: Text Document Processing with the Bayes Formula for Classification Using the Support Vector Machine. IEEE Transaction on Knowledge and Data Engineering 20(9) (2008)
17. `http://www.cs.waikato.ac.nz/ml/weka/`
18. Tan, P.-N., Steinbach, M., Kumar, V.: Introduction to Data Mining. Addison Wesley, Reading (2008)
19. Soucy, P.: Beyond TFIDF weighting for text categorization in the vector space model. In: Proceedings of the 19th International Joint Conference on Artificial Intelligence, pp. 1130–1135 (2005)
20. Dave, K., Lawrence, S., Pennock, D.: Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. In: Proceedings of the 12th International World Wide Web Conference, pp. 519–528 (2003)
21. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn