Fake Filtering in Advertising Restaurant Blogs Using Machine Learning Methods

Jae Young Chang[1], **Han-joon Kim**[2], **and Jong Hoon Chun**[3]
[1]School of Computer Engineering, Hansung University, Seoul, Korea
[2]School of Electrical and Computer Engineering, University of Seoul, Seoul, Korea
[3]School of Software Convergence, Myongji University, Seoul, Korea

Abstract -*Nowadays, users who intend to choose a restaurant based on information provided by blogs are increasing significantly. However, most of blogs are unreliable since restaurant blogs are occupied by advertising postings written by 'power bloggers'. Thus, to ensure the reliability of blogs, it is essential to filter out fake or exaggerated advertising blogs. In this paper, we propose a way of distinguishing the advertising blogs with a machine learning-based classification technique. In the proposed method, we first collect advertising restaurant blogs, and then identify significant features which are commonly found in these blogs. The fake filtering is achieved by developing classification models based on naïve Bayes and neural networks learning algorithms.*

Keywords: Fake Filtering, Machine Learning, Blogs, Classification

1 Introduction

The rapid spread of social media services with the coming of the Web 2.0 era has become the basis of various forums through which the public can express their subjective opinions about the general community. Blogs, which are offered by portal sites, are owned by many users because they provide a platform through which authors can share their knowledge or experiences with their readers. Knowledge from various fields is shared on blogs, and blogs related to information on dining out take up the majority of these. Moreover, many marketing studies are being conducted through these restaurant blogs [1, 2]. This shows that blogs are a significant factor in the selection of a restaurant.

Restaurant blogs involve the author visiting a restaurant and delivering information on their subjective or objective experience to their readers. Readers refer to reviews on restaurants recommended by blogs before selecting a restaurant. However, an excess of various advertisement blogs on the internet focus on content that benefits the company that commissioned the advertisement, instead of offering objective information to the readers. As a result, distorted information is given to readers, and credibility is lost. Because blogs that offer restaurant information are recognized as a type of marketing, portal sites that operate blogs allow the posting of restaurant blogs under the condition that they clearly state that it is an advertisement blog. However, most advertisement blogs contain exaggerated content under the guise that the author actually had such experiences. Therefore, filtering out blogs that contain honest reviews among these blogs is an important factor for offering accurate information to users.

A technique that is similar to advertisement blog filtering is the spam mail filtering technique [3, 4]. Because much research on this technique has already been conducted, it is being used in most mail servers. However, the advertisement blog filtering technique is difficult to implement, unlike spam mail filtering. This is because it is easy to distinguish whether an e-mail is spam or not by looking at the pattern or word distribution in the e-mail, but advertisement blogs are written as if it is not an advertisement. Despite the recent increase in need for this technology, there are still not sufficient studies on this topic [5–7].

This paper proposes a technique that filters advertisement blogs that act like honest reviews (non-advertisement) among restaurant blogs. Filtering uses the machine learning-based automatic classification algorithm [8]. High-quality learning data is required to apply automatic classification. In this study, typical features of advertisement blogs were manually investigated, and this information was used to collect blogs that would be used as learning data. To express the features of advertisement blogs in a quantitative way, various pre-processing techniques were used, such as stopword removal, grammar inspection, and original form conjugation. Based on the features of blogs with review characteristics, many subjective expressions are contained in posts, so an additional sentiment analysis [8] must be conducted to use the results as features that should be distinguished in advertisement blogs. The features of advertisement blogs are used in the classification algorithm that uses learning data, and in order to determine which features to use, a correlation analysis is conducted and various feature combinations are configured using the results of the analysis. The accuracy of classifications is tested by using the classification algorithm that was configured through this combination of features; then, the final algorithm and feature combination are explored. This study used the naïve Bayes classification and the neural network learning algorithm for the automatic classification [8].

2 Data collection

As pointed out in [7], it is extremely difficult to read review blogs and manually assess whether or not the blog post is an advertisement post. Fortunately, if the blog post is an officially sponsored advertisement post, this fact is frankly stated in the post. Therefore, blog posts that clearly state sponsorship were manually collected and their common features were deduced. The results showed that most of the blog post repeated the business name several times in the post, the phrase "tasty restaurant" was also frequently written, and the detailed "address" of the business was included. A considerable number of posts also included the phrase "tasty restaurant" in the title of the post. Therefore, these evaluation criteria that determined whether a blog post contains advertisements are defined in the second row of Table 1.

Conversely, non-advertisement blog posts were collected based on subjective judgments. In order to reduce any distortions in the collection criteria, blog posts that were clearly determined as a non-advertisement post from several experimenters were collected. The evaluation criteria of advertisement blog posts were applied directly to non-advertisement blog posts that were collected, and the features of the results are shown in the third row of Table 1. In other words, the phrase "tasty restaurant" was not included in most of the titles, the phrase "tasty restaurant" and the "business name" was mentioned one time or less, and most of them did not include a detailed address.

Table 1. Data Collection Criteria

Collection Criteria	Advertisement Blog Features	Non-Advertisement Blog Features
Mention of "Tasty Restaurant" [Title]	Yes	No
Mention of "Tasty Restaurant" [Title]	3 or more times	1 time or less
Mention of Business Name [Content]	4 or more times	1 time or less
Mention of Address [Content]	Yes	No

Since a standard was needed for distinguishing whether or not a given blog was a restaurant blog that should be collected for use as learning data, the names of restaurants were collected from a public data portal (https://data.go.kr). Random business names were selected, and blogs were searched using the selected business name and "tasty restaurant" as the keywords. A total of 66,000 restaurant-related blogs were collected using this method.

Of the blogs that were collected, some were extracted as learning data, and learning data was extracted as advertisement blogs or non-advertisement blogs based on matching three or more criteria from Table I. The collected data included 1,132 advertisement blogs and 2,236 non-advertisement blogs for a total of 3,368 blogs.

3 Feature Exploration and Classification

This section explains the process of exploring blog features to define the independent variables of the classification algorithm. First, candidate features are defined; then, features are classified according to the degree of correlation through a correlation analysis between these candidate features and the target variable (advertisement or non-advertisement).

3.1 Definition of Features

In order to extract advertisement blogs by using the classification algorithm, the features to be used for classification must first be defined. In this study, features that are expected to show a difference between advertisement and non-advertisement blogs were defined, and they were generally divided into features regarding blog format and features regarding emotional expressions. Features regarding blog format included certain words, expressions, or forms that appear in the blog, and features regarding emotional expressions included those that were defined according to the type or distribution of emotional words that appear in the blog. The list of features that were defined through blog composition is shown in Table 2.

The top 4 features in Table 2 were the same as the learning data collection criteria in Table 1. The features of the classification model were also included as they are the most important criteria for classifying these into advertisement and non-advertisement blogs in this study. Other features were also defined such as blog post word count, day of the week when the blog was written, length of the blog post, whether or not a map showing the business location was included, number of images, whether or not copyright was displayed, and whether or not the business's phone number was included.

Table 2. Features Defined According to Blog Format

Variable	Variable Description
Tmatzip	Mention of "Tasty Restaurant" Keyword in Title
juso	Mention of Address Keyword
keyword_count	Mention of Keyword (Business Name)
matzip_count	Mention of "Tasty Restaurant" Keyword
word_count	Blog Post Word Count
day	Day of the Week When Post Was Written
content_length	Post Length
map	Map
image_count	Image Count
right	Copyright Display
phone	Phone Number Display

Next, features according to emotional expressions were defined. Since restaurant blogs are written as reviews, most of them included many emotional expressions. Therefore, the difference in emotional expressions were expected to play an important role in categorizing advertisement and non-advertisement blogs. The Open Hangul API was used for the sentiment analysis, which evaluates whether or not words that appear in the blog are positive or negative expressions. When a given word is determined as positive/negative/neutral, the

probability is also shown quantitatively. The probability is shown through a score of 0 to 100, with a higher probability as the score was closer to 100. The features from the emotional expressions that were defined in this paper using this method are shown in Table 3.

In Table 3, the POScore shows that the overall content of the blog is positive. If the sum of all scores of words that are determined to be positive is greater than the sum of scores of words that are determined to be negative, the score is 1. Otherwise, the score is 0. The NAScore gives scores in the opposite way. Therefore, the POScore and NAScore determine the polarity of overall emotion in the blog and have a score of either 0 or 1. On the other hand, CNOScore, CPOScore, and CNAScore are the sum of scores regarding neutral, positive, and negative words, respectively, and have varying quantitative values according to their degree for each blog post, unlike POScore and NAScore.

Table 3. Features Defined According to Emotional Expressions

Variable	Variable Description
POScore	1 if positive score value is greater than negative score value; 0 if otherwise.
NAScore	1 if negative score value is greater than positive score value; 0 if otherwise.
CNOScore	Sum of neutral word scores
CPOScore	Sum of neutral positive scores
CNAScore	Sum of neutral negative scores

3.2 Correlation Analysis

In this study, a correlation analysis was conducted in order to assess the degree of correlation between the features stated in Section 4.1 with the classification of advertisement and non-advertisement blogs. The analysis results were used to determine the combination of independent variables to be used in automatic classification. The Pearson Correlation Coefficient was used for the correlation analysis, and non-advertisement blogs were given a value of 0 and advertisement blogs were given a value of 1. The results of the correlation analysis are shown in Table 4. The degree of correlation in this table was determined according to the correlation coefficient (r) as follows, and this category referenced the correlation analysis page of Wikipedia Korea.

- If $-1.0 < r < -0.7$, then a strong negative correlation (SN).
- If $-0.7 < r < -0.3$, then a clear negative correlation (CN).
- If $-0.3 < r < -0.1$, then a weak negative correlation (WN).
- If $-0.1 < r < 0.1$, then a negligible correlation (IGN)
- If $0.1 < r < 0.3$, then a weak positive correlation (WP).
- If $0.3 < r < 0.7$, then a clear positive correlation (CP).
- If $0.7 < r < 1.0$, then a strong positive correlation (SP).

As shown in Table 4, the top 4 variables, which are the learning data collection criteria, all have a strong or clear positive correlation. Because they were the first criteria for categorizing advertisement and non-advertisement blogs, this is understood as obvious results. From the remaining features, mention of the phone number (phone) had a strong correlation, and post length (content_length), inclusion of a map (map), and the sum of emotion scores regarding neutral or positive words (CNOScore, CPOScore) had a clear correlation. Conversely, POScore or NAScore, which only showed the simple polarity (0 or 1) regarding the blog post's overall emotion had a weak correlation. This implies that there are almost no blogs that express negative emotions regarding restaurants.

Table 4. Results of Correlation Analysis and Feature Classification

Variable	Correlation Coefficient	Degree of Correlation
Tmatzip	1	SP
juso	1	SP
keyword_count	0.818	SP
matzip_count	0.696	CP
word_count	0.500	CP
day	0.004	IGN
content_length	0.345	CP
map	0.435	CP
image_count	0.209	WP
right	0.134	WP
phone	0.770	SP
POScore	0.203	WP
NAScore	−0.118	WN
CNOScore	0.537	CP
CPOScore	0.479	CP
CNAScore	0.287	WP

4 Automatic classification and evaluation

This study used the naïve Bayes and neural networks-based classification algorithm as a means to extract non-advertisement blogs. The independent variables were divided into seven different combinations as shown in Table 5. This combination of independent variables was determined according to the degree of correlation in Table 4. According to Table 4, there were almost no variables with a negative correlation. Therefore, only the degree of correlation was considered regardless of positive or negative correlation.

Table 5. Definition of Independence Variables

Combination of Independent Variables	Description
All	All Variables in Table 4
Basis	Learning Data Collection Criteria (4 types)
C	Clear Correlation
S	Strong Correlation
W+C	Weak Correlation and Clear Correlation
S+C	Strong Correlation and Clear Correlation
W+S	Weak Correlation and Strong Correlation

For advertisement and non-advertisement blogs, we prepared 200 blog data, respectively. From the blogs that were collected automatically for learning regarding advertisement/non-advertisement blogs, 100 of each were extracted at random and used as test data.

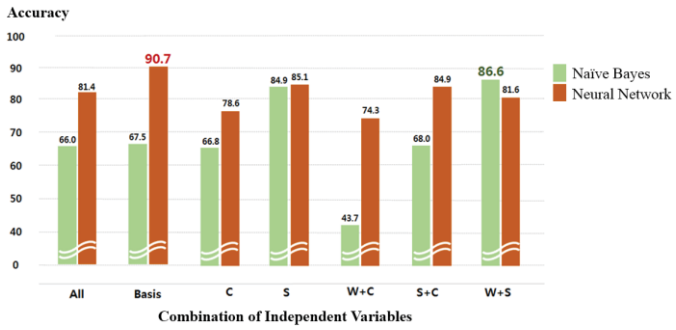

Fig. 1 Comparison of classification accuracy

The test results are shown in Fig. 1. First, in the naïve Bayes classification, the accuracy of *S* and *W+S* were 84.9% and 86.6%, respectively. If a variable with a strong correlation was included, this would show good classification. Conversely, *W+C* had the lowest performance at 43.7%. The remaining combinations that included *Basis*, which is the combination of data collection criteria, had an accuracy of around 60%. *Basis* includes the most important variables that configured learning data. *C* or *S+C* is a combination of variables with a high correlation, but this had a lower accuracy than *W+S*, which included a weak correlation.

The neural network classification showed a considerably better performance than the naïve Bayes classification. Regardless of which features configured the combinations, there was a relatively even accuracy of around 70–80%. The model built through Basis had the highest accuracy at 90.7%. Based on these results, the neural network is more appropriate than the naïve Bayes as a classification technique for extracting advertisement blogs, and configuring independent variables with features that have a strong or clear correlation were shown to result in a better performance.

5 Conclusions

This study proposed a method of filtering advertisement blogs that exaggerate reviews on restaurants. For the proposed method, an automatic classification method was used, and this involved the process of collecting blogs, pre-processing, correlation analysis, feature selection, classification, and evaluation. To collect advertisement and non-advertisement blogs, advertisement blogs were manually selected, and their features were compared to non-advertisement blogs and extracted accordingly. A sentiment analysis was also conducted, and the results were added to the features. The typical features of advertisement blogs were analyzed quantitatively through a correlation analysis on the finalized features. The classification method was applied through various combinations of the results. Combinations with features that have a higher correlation were found to have higher classification accuracy, and the neural network algorithm had a higher accuracy than the naïve Bayes algorithm. In the future, we plan to design a way of evolving the advertisement blog filtering technique according to the situation instead of having fixed features.

6 Acknowledgements

This work was supported by Small and Medium Business Administration (SMBA) of Korea in 2017 (No: S2428187), and was also supported by the Science and Technology Expert Technology Supporters program through SMBA of Korea in 2017 (No: C0440541).

7 References

[1] J. W. Kim, and I. Y.Kim, "How the characteristics of the food-blog marketing effect to purchasing intension with the mediation effect of trust," Korean Journal of Tourism Research, Vol. 30, No. 5, pp. 85-105, 2015.

[2] J. K. Kjm, H. K. Kim, and S. Y. Park, "A study on blog user's response to blog marketing," information Systems Review, Vol. 11, No. 3, pp.1-17, 2009.

[3] E. Blanzieri, and A. Bryl, "A survey of learning-based techniques of email spam filtering," Artificial Intelligence Review, Vol. 29, No. 1, pp. 63-92, 2008.

[4] G. Cormack, "Email Spam Filtering: A Systematic Review," Foundations and Trends in Information Retrieval, Vol. 1, No. 4, pp. 335-455, 2007.

[5] I. S. Park, H. H. Kang, and S. J. You, "Classification of Advertising Spam Reviews," Proceedings of the 22nd annual conference on Human & Cognitive Language Technology, Korea Information Science Society, pp. 186-190, 2010.

[6] H. Ahn, B. J. Park, "Extracting similar advertising review for Opinion Mining," Proceedings of the annual conference on Electronics and Information Engineering, pp.1593-1596, 2014.

[7] N. Jindal, and B. Liu, "Opinion Spam and Analysis," Proceedings of WSDM, pp. 219-229, 2008.

[8] J. Chang, and I. Kim, "An Experimental Evaluation of Short Opinion Document Classification Using A Word Pattern Frequency," Journal of the Institute of Internet, Broadcasting and Communication, Vol. 12, No. 5, pp. 243-253, 2012.

[9] A. Mukherjee, V. Venkataraman, B. Liu, and N. S. Glance, "What yelp fake review filter might be doing?," Proceedings of International AAAI Conference on Web and Social Media, 2013.