International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 10, No.1 (Jan-2021)
http://dx.doi.org/10.12785/ijcds/100104

Automated Product Review Collection and Opinion Analysis Methods for Efficient Business Analysis

Ill Chul Doo[1], Hyun Duck Shin[2] and Mee Hwa Park[3]

[1] *Hankuk University of Foreign Studies, Korea*
[2] *Convergence Institute, Hansung University, Seoul, Korea*
[3] *Software Focused University, Korea*

Received 5 Mar. 2020, Revised 9 Jul. 2020, Accepted 5 Aug. 2020, Published 1 Jan. 2021

Abstract: In this paper, we propose a business analysis method that automatically collects review information related to a specific product by using a web crawler, analyzes the customer's emotional response to the product, and supports marketing activities. In order to process data collected from web pages and social networks into information that can be used for marketing, several levels of data processing and text mining techniques are needed. Although various studies have been carried out for this purpose, the data collected through lots of effort and cost contain more extensive information than needed for marketing. So the usefulness of the information obtained through data processing and analysis is not really good. In this paper, we propose a system that automatically receives data from interested sites, interested fields, interested keywords, and target period information from a user and crawls the data. In addition, the reviews that can be used only in marketing can be selected to judge whether or not they are positive, thus it improves the accuracy of the analysis. For the experiment, we collected the reviews of the books sold by specific publishers over the past three years and conducted reputation analysis. The proposed classifier distinguishes through supervisor whether the data collected is a proper review or not. The accuracy of the review classifier is 98.7%. The reputation analyzer judges whether the review is positive or negative with the 86.1% accuracy. The results of this study can be used directly in various industries. And we plan to develop the reputation analyzer, improving the accuracy and extracting the reputation factors affecting customers.

Keywords: Business Analysis, Opinion Mining, Customer Review Analysis, Automatic Data Collection

INTRODUCTION

In this paper, we propose a system to support marketing that can automatically collect information related to a specific keyword using a crawler and analyze the opinion of the text containing the keyword. The proposed system automatically classifies the crawled texts into customer review data and general data, and then judges if it is positive opinion or negative to provide with useful data for marketing. With the proliferation of smartphones, there are increasing numbers of consumers interacting with others with various opinions on products through Web pages and SNS, and consumers searching for product review information just before purchasing products. The consumers share their opinions naturally via social media such as Twitter, blogs, and Facebook, as well as the site where they purchased the product, and receive practical help when purchasing products. As a result, web pages and social media are attracting attention as one of the best marketing indicators that can understand consumers best in terms of the data pool produced by consumers voluntarily, rather than a simple communication tool.

The sites such as Amazon [1] and Review Center [2] are increasing the probability to attract the consumers to purchase by effectively managing user reviews about the product as well as information about the product itself. Also, the companies in various fields such as movies [3], lodging companies [4], and bookstores [5] analyze customers' product review data and offer the customers' evaluation scores to attract customers and use them for marketing [6-7].

The data created by the user includes their opinions, feelings and thoughts. The natural language processing technique for analyzing such subjective data is called opinion mining or sentiment analysis. For marketers, emotional analysis is an indispensable tool for business success. In order to acquire information that meets marketing targets through emotional analysis, it is necessary to collect and analyze data with programming technology, and marketing techniques. However, it is a real challenge for marketing professionals to collect data and analyze on their own.

E-mail address: dic@hufs.ac.kr, henry@hansung.ac.kr, maia.mhpark@gmail.com

There are Business intelligence (BI) tools provided by vendors to support marketer's analysis and business decision making. The BI tools are application software that collect and process large amounts of data in one's organization. They are not as flexible as business analytics tools, but they often offer the way to gather data to find information through queries, report, and dashboards, and help prepare data for analysis so that you can create any reports or data visualization. They process large amounts of unstructured data from internal and external systems, including books, journals, documents, health records, images, files, emails, videos and other business sources. However, existing BI tools are not suitable for selecting, collecting customer review data related to products and performing emotional analysis on the Internet and social networks.

In order to compensate for this, many studies related to sentiment analysis have been conducted [8-16], in which users' opinions are automatically extracted and analyzed. They analyze the customer's review data collected from e-commerce sites to judge whether the user's opinions are positive or negative, or to summarize the information using natural language processing techniques, text mining, and statistical analysis. However, external data collected from web pages or SNS contains a lot of information that cannot be used for marketing. In the internal data provided by the e-commerce system, users do not write the only evaluation of the product, but also various contents such as questions about the problem during delivery, how to use the product, daily life or spam text. The information obtained by analyzing such data without filtering is hardly useful. The method proposed in this paper automatically collects data by user's inputting sites, fields, and keywords they are interested in during targeted period. Then, we conduct a reputation analysis on the product by selecting only product reviews that can be used for marketing. As a result, we can improve the accuracy and utilization of the univariate analysis of the product.

This study is composed as follows. The Chapter 2 introduces related research on opinion mining. The Chapter 3 explains the structure and emotional analysis of intelligent product review collection and analysis system. Experimental data and results are presented in Chapter 4, and conclusions and future studies are discussed in Chapter 5.

RELATED RESEARCH

The opinion mining refers to a technique for extracting or classifying opinions expressed in various contents described by users in texts, such as online news and social media comments with various techniques of emotional analysis. That is, it is a technique of determining whether a sentiment of a person who created the document is 'positive' or 'negative' in a text document in which the context is not defined. The opinion mining technology effectively extracts and summarizes users' product reviews, allowing potential consumers to easily browse and search for various opinions on products without looking for all product reviews. Emotional analysis is used to analyze product reviews of online shopping malls. In other words, emotional analysis results are used for marketing in various industries such as movies [3], accommodation [4], books [5-7], and soap operas. [8]

Emotional analysis can be divided into linguistic method and semantic analysis method. Linguistic methods use natural language processing algorithms and machine learning. It uses the functional and emotional information about the parts of speech or vocabulary of text data to extract the features representing the vocabulary function and judge the positive and negative sensibility of the text [9]. That is, when the expert defines the vocabulary patterns used as opinions, the vocabulary corresponding to the pattern is selected as the target vocabulary. Then, if there is a vocabulary that expresses or describes the opinion, the vocabulary is extracted. And then it judges if the target vocabulary is positive or negative with various algorithms. Esuli, A., and Sebastiani, F[10] used a machine learning algorithm to learn the classifier with the document containing opinion and the two extreme information and then determined the semantic extreme of the new document.

A method for extracting vocabularies based on the frequency of vocabularies included in a document, numerical values such as TF-IDF, or statistical techniques has also been attempted [11-13]. Xiaowen Ding, and Bing Lui [13] attempted emotional analysis using statistical methods for language rules such as the relationship between sentence structure and sentences, and the pattern information of sentence components. There is also a study on the semantic analysis method in which the language rules such as the relationship between the sentence structure and the sentence. And the pattern information of the sentence component are classified using a language dictionary such as WordNet [14]. These studies use emotional dictionaries that tag positive, negative, and neutral at the level of the vocabulary or language for the contents of individual languages. In semantic analysis method, collected linguistic resources and its emotional classification information (called emotional dictionary or emotional corpus), are very important factors [15]. If the sensitivity of the emotional vocabulary is not clearly defined, the accuracy of the emotional difference algorithm is lowered. The studies using emotional analysis based on English texts have been conducted with SentiWordNet [16]. They use WordNet to determine the positive or negative meaning of a vocabulary and then apply it to SentiWordNet to quantify the extent of emotion.

Hangul is not easy to analyze emotion because its characteristics are different from English. Moreover,

since the language used on the Internet is not registered in the Korean dictionary, it is important to build a sensitive dictionary specialized for analytic purpose. There are some studies that build emotional dictionaries considering the grammatical and semantic characteristics of Hangul [17-19]. Sukjae Choi and Ohbyung Kwon [17] developed SentiWordNet in Korean by identifying emotional vocabularies included in the sentence and determining the degree of emotion. There are some studies that design emotional dictionaries using Korean antonyms [18] and emoticons, special symbols used in the Internet, and emotional symbols of Korean initials [19]. Liu, B et al [20] proposed a technique for summarizing product reviews using machine learning algorithms and natural language processing techniques. As a result, we have developed a system named Opinion Observer for research. The Opinion Observer offers summarized results by analyzing the opinion information of the review document mainly about the product. After extracting opinion information in a predefined pattern by using natural language processing technique, we judge the both extremes of opinion information by using WordNet. The OPINE [21] system extracts the vocabularies of interest using predefined sentence structure patterns. At this time, the PMI value between the sentence structure and the vocabulary is calculated, and the vocabulary of the opinion is selected and finally the opinion vocabulary associated with the target vocabulary is determined through the parsing. The Red Opal [22] system, developed by Carnegie Mellon University in the United States, generates a summary report using the user's appraisal and score. The Red Opal system displays multidimensional analysis of product reviews and scores, but does not evaluate whether subjective opinions are positive or negative.

Existing studies analyzing product reviews provided by merchandising sites do not reflect user opinions expressed in various forms on the Internet. In other words, it will not reflect various opinions of the products displayed through web pages or social media. Even if various information is collected on the Internet, if the data that does not fit the purpose of analysis is used for analysis without being filtered, the usage of the result will be decreased. Therefore, this study aims to offer useful information for marketing by automatically collecting various data from web pages and social media, and then selecting only the necessary data for analysis. To this end, we propose a review classification method and emotional analysis method that classify the data collected from the source of information to be analyzed into review and non-review data through map learning. In order to verify the superiority of the method proposed in this paper, the review information about the books sold in the past 3 years by A publishing company was collected and emotional analysis was conducted. We also reviewed the impact of review on sales by verifying the availability of the review data as marketing source.

CUSTOMER REVIEW COLLECTION AND OPINION EVALUATION SYSTEM

A total of 30 participants at Information on the Internet can be divided into two types: "fact" and "opinion". The facts refer to the objective information described in relation to the universal phenomenon, and the opinion refers to individual subjective information about a phenomenon. Currently, most search engines on the Internet are focused on the search for facts, then the quantity of opinions is insufficient. However, when searching the Internet, it often happens that individual opinions are needed as well as facts. For example, many people have a significant influence on the choice of people who want to purchase the product by posting subjective opinions on the product on the Internet. The opinions include both advantages and disadvantages that the producer did not know, so the opinions can be used for product research and marketing in a comprehensive way.

In this paper, we propose a system that automatically collects and analyzes opinion data on products on web pages and social networks and provides information that can be used in marketing.

3.1 System configuration

The proposed system consists of 4 modules as follows.

1) Data Collector: A module that collects related data automatically when the user inputs the product information and the period to search, and saves them as individual files

2) Data Classifier: A module that classifies data sentences into review and general data

3) Opinion Analysis: A module that classifies users' opinions as positive or negative and calculates reputation scores

4) Business Analysis: A module that analyzes the impact of reputation scores on products on sales or forecasts sales demand

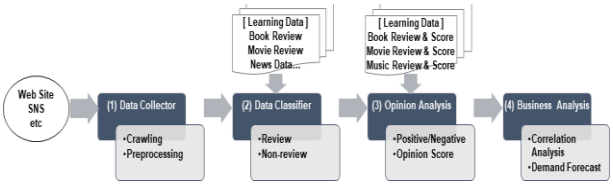

Figure 1. Composition of Customer Review Collection and Opinion Evaluation System.

The proposed system configuration is shown in Fig 1. The proposed system was developed by Python Program. The Data Classifier and the Opinion Analyzer perform supervised learning based on the data. The Data Sorter separates the crawled text into words, and reviews such

as on movie or internet bookstore. The Opinion Analyzer conducts supervised learning using movie review and rating data collected from internet bookstores and judges whether the opinion on the article is positive or negative.

3.2 Collecting and classifying opinions

The review data collection and classification module collect related data automatically when the user inputs the product information and the term to be searched, and divides the data into sentences and stores it as a separate file.

3.2.1 Data Crawling Module

Crawling is a technology that automatically collects documents from a Web site on the Web providing vast amounts of data. The web crawler implemented in this paper consists of 4 steps. In the first step, the target page is determined based on the URL of the target site to be collected. In the second step, the HTML source code of the web page is analyzed to determine the location of the data to be collected. The third step it separates the collected data from the HTML tags. The last step is to save the collected data as a "csv" file.

Figure 2. Process of the Website Crawling.

3.2.2 Data preprocessing

Unlike numerical data, which is structured and directly processed by a computer, text data needs to be converted into a structure that is easy to analyze. In this paper, we classify the reviews and divide the collected data into sentences to identify the emotions included in the review. We used the KoNLPy library to process reviews in Korean. KoNLPy is a Python package for native language processing (NLP). BOW (Bag of Words) was generated by segmenting into sentence units using the split function of KoNLPy and then dividing them into morpheme units with the morphs function.

Figure 3. Process of the Website Crawling.

The bag-of-words model is a simplifying representation used in natural language processing and information retrieval. In this model, a text is represented as the bag of its words, disregarding grammar and even word order but keeping multiplicity. The bag-of-words model is commonly used in methods of document classification where the occurrence of each word is used as a feature for training a classifier.

Bag-of-word model is a technique for determining the type of a document by looking at the frequency of occurrence of words contained in the text. For example, if a document contains many words such as 'exchange rate', 'stock price', 'interest rate', etc., this document will be categorized as documents related to economics. And if the words 'backlighting' or 'exposure' are found in a document, it classifies it as the document about photography. The Bag of Words model classifies the document as a frequency value of words without considering the context or sequence of Corpus, or obtains the similarity between documents. Unlike English, word order is not important in Korean. Therefore, BOW can be useful in Korean text analysis. However, when a feature vector for a word is generated using the BOW, a very large number of words become a Sparse Matrix type in which a column value of a vector is filled with 0, which requires a lot of memory space and a complicated data operation. To improve this, we used the Scikit-Learn Python library.

Scikit-Learn is a simple and efficient tool for data mining and analyzing. The Scikit-Learn TfidfVectorizer class uses the Compressed Sparse Row (CSR) format to resolve spatially matched problems while identifying important words using the TF-IDF technique. TF-IDF (Term Frequency-Inverse Document Frequency) is a weight used in information retrieval and text mining. It means a statistical value that indicates how important a word is in a particular document when there is a document group consisting of several documents. The TF-IDF weight can be used to extract key words of a document, to rank search results in a search engine, or to calculate similarities between documents.

In TF-IDF, TF (term frequency) is a frequency value indicating how many words appear in a document. In general, if the TF is high, the word is likely to be an important word in the document. However, if the word is a commonly used one, it is difficult to consider it to be important in the document. The concept introduced to correct this problem is the IDF (inverse document frequency). IDF is the inverse of DF (document frequency). The higher the weight of the document containing the specific word in the entire document, the lower the IDF value in inverse proportion creates. In practice, the IDF calculates a log by dividing the total number of documents by the number of documents containing the word. Thus, the TF-IDF weight for the final specific word is calculated by multiplying TF and IDF.

Although the word order of Korean words is not important, we used an n-gram model that separates words into one or two words considering the

characteristics of Korean, which often uses adverbs or emotional expressions. In other words, when converting a sentence collected by TfidfVectorizer into a word vector, the value of 'ngram_range' is set to (1,2), and the character is extracted with two tokenized words and two consecutive words. The 'min_df' value was set to 3 in order to exclude meaningless words, frequencies, and low-level words, and the 'max_df' value was set to 0.9 in order to exclude words with too high frequency.

3.2.3 Data classification

All of the data collected on web pages and social networking sites may not be user comments on products. For example, it is necessary to classify various types of articles, such as questions, advertisements, and news, rather than review. So the machine learning methods were used in this paper. Machine learning is tried to learn computers like humans and to form rules on their own. Machine learning can be divided into Supervised learning, Unsupervised learning, and Semi-supervised learning. Supervised Learning is a way in which a person gives data to a computer with a label (y) for each input (x) as a teacher and the computer learns it. It has the advantage of using highly accurate data because it is directly intervened by people. Instead, there is a labor cost problem because people have to label them directly, and one more problem that the amount of data that can be obtained is small. Unsupervised Learning is the process by which a computer learns about data that is not self-labeled. In other words, you learn using only x without y. There is a close relationship between statistics clustering and estimation of distribution by solving problems without correct answers. Semi-supervised learning is to learn from both labeled and unlabeled data, often with a problem of supplementing a large number of unlabeled data with a small amount of labeled data. In this study, the review was categorized as a map learning method using data for learning. In other words, review data were labeled as 1 for learning and non-review data as 0 for learning collected instead of reviews. The review data consisted of data crawled from movie reviews and internet bookstore reviews on the portal site. The learning data of the review classification model is about 230,000 sentences and the test data are 70,000.

The algorithm used to determine if the data is review or not is a logistic regression model. Logistic regression is a probability model proposed by D.R.Cox in 1958[1], a statistical technique used to predict the probability of an event using linear combinations of independent variables. The purpose of the logistic regression is to use the relationship between dependent and independent variables as a concrete function in the future prediction model, as in the general regression analysis. This is similar to the linear regression analysis in that it describes the dependent variable as a linear combination of independent variables. However, logistic regression can be seen as a sort of classification technique because, unlike linear regression, the dependent variables are targeted at categorical data and, given input data, the results of that data are divided into specific classifications.

In logistic regression, the active function that determines whether the input function is weighted or not is the Odds. Odds are the probability of something succeeding. If the probability of a particular event occurring is p , The odds ratio is $\frac{p}{1-p}$. That is, the multiplication factor is defined as the ratio of the probability that a certain event will occur and the probability that it will not occur. If a function having the log value of the multiplication ratio as a function value is defined, $f(x) = \log \frac{p}{1-p}$. If you convert this formula into an exponential formula, it is $p = \log \frac{1}{1+e^{-x}}$. The larger the output value of the active function used in the logistic regression model is, the more converges to a value of 1. The smaller the output value is, the more converges to 0. Generally, it is used as a model to classify as false if it is 0.5 or more, and false if it is less than 0.5.

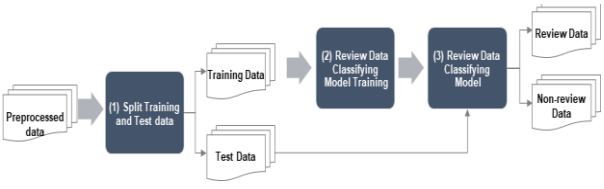

Figure 4. Process of the Data Classification.

As a result of reviewing and classifying the study by using 350,000 sentences gathered for review classification, the accuracy of the review classification model was 98.7%. The test results in detail are discussed in Chapter 4.

3.3 Opinion Analysis

Text data classified as reviewed text goes through a process of opinion analysis to determine whether a user's opinion is positive or negative. The opinion analysis modules are initially constructed using logistic regression models because they are a single item classification process which is similar to review classification modules. 150,000 learning data and 50,000 test data were used to determine whether the model was positive or negative. A total of 200,000 data are collected from web pages, users' posts, rating movies or products.

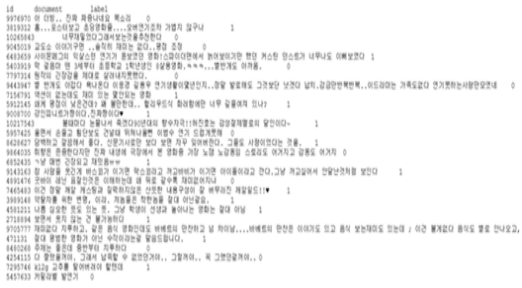

Figure 5. Sample of learning data used in opinion analysis.

The k-Fold Cross Validation method is used to prevent the review classifier and the opinion analysis model from overfitting the learning data and to improve the accuracy of the model by appropriately selecting Hyperparameter values in machine learning. If the dataset is separated into Train Set and Validation Set, and then the Train Set is learned and then the actual model is tested with the Validation Set, the accuracy may be low. This case is called overfitting, which is very well suited for a particular dataset but not for another dataset. Cross-validation is one way to avoid this problem. The k-Fold Cross Validation divides the Train Dataset evenly into k groups (called Folds) and assigns them to (k - 1) Test Folds and one Validation Fold. For each verification, Test Fold is specified differently and total k times verification is performed to measure performance. After performing k-times verification, the average value of k hyperparameter values is obtained, and this value is used as the hyperparameters of the final model.

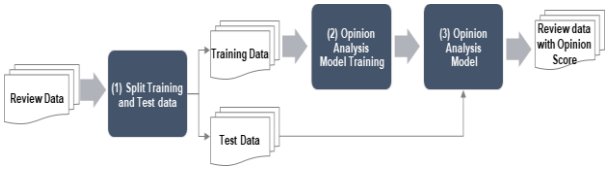

Figure 6. Process of the Opinion Analysis.

RESULTS OF COLLECTION AND ANALYSIS OF BOOK OPINIONS

The automatic review classification and emotional analysis model proposed in this study was applied to the book to validate the usefulness for marketing. We finally analyzed the correlation between the users reviews on the book and the sales status data over the three years (2014-2016) provided by Company A

4.1Data Collection and Analysis

The data collected in this study for experimentation are divided into two main categories. First, it is learning data of review classification model and emotion classification model. The second data is book-by-book review data offered by Company.

In order to study the review classification model, we collected articles as reviews and non-reviews respectively. The review article is 199,992 on movie in portal sites and textbook reviews on Internet bookstore sites. Non-review article is 159,643 by selecting text that does not include the user's opinion. To create the emotional analysis module, we secured 99,996 positive reviews and 99,996 negative reviews based on a total of 199,992 data

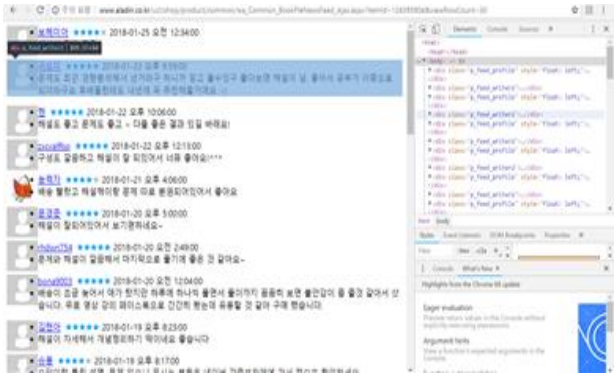

Figure 7. Screen to crawl internet bookstore reviews with ratings.

The review classification module, which collects user reviews of books and then selects only fine reviewers, and the emotional analysis model, which determines positivity/negative, are based on the machine learning, the logistic regression model. The both models divided into a train set and a test set in a ratio of 3:1. As a result of the trial, the accuracy of the review classification model was 98.7%, and the accuracy of the emotional analysis classification model was 86.1%.

Table 1. Data collection results

	Site A	Site B	Site C	Site B
Crawling Data	74,271	33,739	89,837	197,847
Review Sentence	164,069	775,623	236,840	1,176,532

To use the system in this thesis, users can select the website they want to search for and enter the name and duration of the book. That is, if a marketers choose the period, library name and site they wants to analyze, the data is automatically collected, pre-processed, classified into a text rather than a review, and the emotional analysis results of the review are saved as a file. The articles about the 425 books that A company have sold for three years are collected on the portal website and the book sales site. And the classification results are followed.

4.2 Marketing Usability Evaluation

In order to analyze whether the review articles on the books classified above affect the sales of the book or not, the dependent and independent variables were defined below and correlation coefficient analysis was performed.

Table 2. Marketing Usability Evaluation

Variable type	Variable name	Description
Dependent variables	Sales assistant	Sales of sales by book from 1 January 2014 to December 31 st 2016
Independent variables	SiteA_ Review_ Number	Number of review sentences collected by SiteA for this book
	SiteB_ Review_ Number	Number of review sentences collected by SiteB for this book
	SiteC_ Review_ Number	Number of review sentences collected by SiteC for this book
	Interest/fraud	The index, which indicates whether reviews of the books are positive, is positive. The positive index, which is close to 1 to 1, is positive

Correlation is a number that represents the degree of correlation between two variables, X and Y, and refers to the relationship between two variables as one of the two increases or decreases as the other. Correlation coefficients can be represented by numerical values which is different from causal relations, and it is judged that a strong correlation is found when the absolute value is closer to 1, a positive correlation is found when the value is larger than 0, and a negative correlation is found when the absolute value is smaller than 0. The absolute value of the correlation coefficient represents the correlation between the two variables and generally divides the degree of correlation on a 0.5 basis.

Correlation analysis shows that there is a strong positive correlation between the number of reviewers collected from three sites and the number of copies sold. However, the coefficient of correlation between the positive and negative indices and the circulation of the sales was –0.038 with little correlation. This reflects the fact that customer interest is more likely to affect sales than book review.

Table 3. Correlation analysis among customers

	SiteA_ Review_ Number	SiteB_ Review_ Number	SiteC_ Review_ Number	Interest/fraud
Correlation	0.714	.536	0.737	-0.038
Results	A strong relationship	A positive relationship	A strong relationship	No relationship

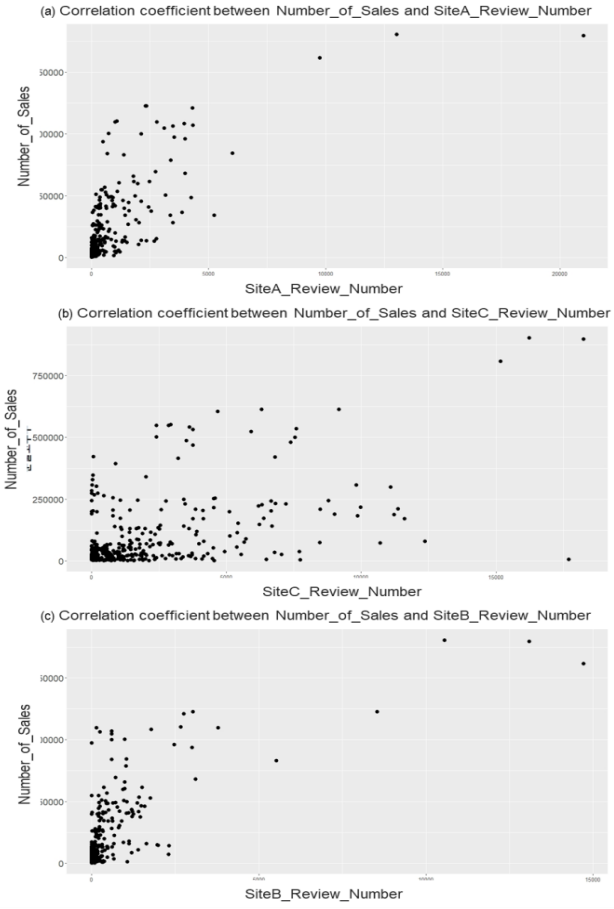

Figure 8. Correlation graph between customer reviews and sales.

Experiments have confirmed the possibility of using the proposed method for marketing, but it is necessary to examine more about the experimental result that the positivity or negativity does not directly affect the sales of the product. That is, it is necessary to analyze whether the patterns are specific only to the books or common to the general products. Further research will be conducted by expanding the scope of product and analyzing more data collection.

CONCLUSION AND FUTURE RESEARCH

In this paper, we proposed a business analysis method that uses a web crawler to automatically collect review information related to a particular product, analyze the customer's emotion on the product, and use it in marketing activities.

The proposed system provides with a data collection function that automatically collects and pre-processes data by inputting the areas of interest, keywords of interest, and target period to collect data for specific products. In addition, the accuracy and utilization of the analysis was improved through review classifiers and emotion analyzers that perform reputation analysis of

products by selecting only product reviews from the collected data that can be used for marketing.

A review and reputation analysis of books sold by some publishers over the past three years has been conducted to find the usefulness for marketing. Experimental results show that the review classifier has 98.7% accuracy and the reputation analyzer classifies the customer's opinion with 86.1% accuracy.

The results of the review sentence appraisal analysis are effective in marketing activities such as forecasting or recommending sales of specific products. In order to determine whether the review information for a book is relevant to the sales of the book, we analyzed the correlation between the results of the emotional analysis and the sales of the book. As a result, it was found that the case of review classification had an average correlation coefficient of 2.12 times higher than that of no review classification.

Using the system proposed in this paper, marketers can quickly understand customer response information about a product without collecting and reading a huge amount of review data. The results of this study can be used directly in various industries. However, we will improve the accuracy of the reputation analyzer by collecting learning data from various sources such as cafes and blogs in order to enhance the usage of the proposed system. And we plan to expand the system to provide more information on important factors that affect customer reputation.

ACKNOWLEDGMENT

This research project is supported by Hansung University's internal research fund.

This research was supported by Hankuk University of Foreign Studies Research Fund (of 2021)

REFERENCES

[1] Amazon [Online]. Available: http://www.amazon.com.

[2] Review Centre [Online]. Available:

[3] https://www.reviewcentre.com/.

[4] Extreme Movie [Online]. Available:

[5] http://www.extmovie.maxmovie.com

[6] skyscanner [Online]. Available: http://www.skyscanner.co.kr

[7] KYOBO Book[Online]. Available: https://www.kyobobook.co.kr

[8] Aladin [Online]. Available: http://www.aladin.co.kr

[9] YES24 [Online]. Available: http://www.yes24.com

[10] Hyunwoo Hwangbo, Jonghyuk Kim, "A Study on Analyzing Sentiments on Movie Reviews by Multi-Level Sentiment Classifier", The e-Business Studies, Vol. 17, No. 6, pp.87-99. 20169. Ding, X., Liu, B., and Yu, P. S., "A holistic lexicon-based approach to opinion mining,"In Proceedings of the international conference on Web search and web data mining, pp. 231-240, 2008.

[11] Esuli, A. and Sebastiani, F., "Determining Term Subjectivity and Term Orientation for Opinion Mining," In Proceedings of 11th conference of the European chapter of the Association for Computational Linguistics : EACL, pp. 193-200, 2006.

[12] Scaffidi, C., Bierhoff, K., Chang, E., Felker, M., Ng, H., and Jin, C., "Red Opal : Product-Feature Scoring from Reviews," In Proceedings of the 8th ACM conference on Electronic Commerce, pp. 182-191, 2007.

[13] Jin, W., Ho, H., and Srihari, R., "Opinion-Miner : a novel machine learning system for web opinion mining and extraction," In Proceedings of the 15th ACM SIGKDD international conference on Knowledge Discovery and Data mining, pp. 1195-1204, 2009.

[14] Xiaowen Ding, and Bing Lui, "The Utility of Linguistic Rules in Opinion Mining," SIGIR 2007, pp. 811-812.

[15] Miller, G., Beckwith, R., Fellbaum, C.,Gross, D., and Miler, K., "Introduction to WordNet : An on-line lexical database," International Journal of Lexicography, pp. 235-244, 1990.

[16] Hu, M. and Liu, B., "Mining and summarizing customer reviews," In Proceedings of the 10th ACM SIGKDD international conference on Knowledge Discovery and Data mining, pp. 168-177, 2004.

[17] Denecke, K., "Using SentiWordNet for Multilingual Sentiment Analysis," In Proceedings of the International Conference on Data Engineering : ICDE, Workshop on Data Engineering for Blogs,Social Media, and Web 2.0, 2008.

[18] Sukjae Choi, Ohbyung Kwon. "The Study of Developing Korean SentiWordNet for Big Data Analytics - Focusing on Anger Emotion". The Journal of Society for e-Business Studies Vol.19, No.4, pp.1-19, 2014

[19] Ji-Hoon Seo, Hye-Jin Jo, and Jin-Tak Choi. "Design for Opinion Dictionary of Emotion Applying Rules for Antonym of the Korean Gramma" Journal of KIIT. Vol. 13, No. 2, pp. 109-117, Feb. 28, 2015

[20] Kyoungae Jang, Sanghyun Park, Woo-Je Kim, "Automatic Construction of a Negative/positive Corpus and Emotional Classification using the Internet Emotional Sign", Journal of KIISE, Vol. 42, No. 4, pp. 512-521, 2015.

[21] Liu, B., Hu, M., and Cheng, J., "Opinion observer : analyzing and comparing opinions on the Web," Proceedings of the 14th international conference on WWW, pp. 10-14, 2005.

[22] Popescu, A. and Etzioni, O., "OPINE : Extracting product features and opinions from reviews," In Proceedings of the conference on Human Language Technology/Empirical Methods in Natural Language Processing : HLT/EMNLP, pp. 339-346, 2005.

[23] Scaffidi, C., Bierhoff, K., Chang, E., Felker, M., Ng, H., and Jin, C., "Red Opal : Product-Feature Scoring from Reviews," In Proceedings of the 8th ACM conference on Electronic Commerce, pp. 182-191, 2007.

FIGURE CAPTIONS

Fig. 1. Composition of Customer Review Collection and Opinion Evaluation System.

Fig. 2. Process of the Website Crawling.

Fig. 3. Process of the Website Crawling.

Fig. 4. Process of the Data Classification.

Fig. 5. Sample of learning data used in opinion analysis.

Fig. 6. Process of the Opinion Analysis.

Fig. 7. Screen to crawl internet bookstore reviews with ratings.

Fig. 8. Correlation graph between customer reviews and sales.

Ill Chul Doo Hankuk University of Foreign Studies. His Ph.D. degree in Digital Culture & Contents from Hanyang University, Korea. He is currently a Professor at Hankuk University, Korea. His interests are mobile contents, and cultural industry, and cultural technology.

Hyun Duck Shin Convergence Institute, Hansung University, Seoul, Korea. His Ph.D. in cultural content studies at Hanyang University, Korea. and He is currently a professor at Hansung University in Korea. Among the areas of interest are the cultural industry, virtual reality and start-up.

Mee Hwa Park Software Focused University, Korea. Areas of Interest: Big Data Analytics, IoT Service, Multimedia Database.