A Study of Analyzing on Online Game Reviews using a Data Mining Approach: STEAM Community Data

Ha-Na Kang, Hye-Ryeon Yong, and Hyun-Seok Hwang

***Abstract*—As the internet environment evolves and new media emerge, consumers start to share their opinions and reviews of products on the web. There is also growing demands for analyzing such online reviews and identifying consumers' true minds to meet these emerging trends while a large number of studies have been made on online reviews in a wide range of academic fields including marketing, MIS and computer science. However there has been little research conducted on video game industries dealing with typical experiential products. Thus, this study was intended to analyze community data in games domain available on STEAM, a world-wide game platform, using a data mining approach. Several machine learning techniques such as Classification and Regression Tree (CART), Artificial Neural Network (ANN) were applied to community data collected from STEAM games to analyze factors that have impact on helpfulness of game reviews. We also conduct sentiment analysis of review comments to mashup sentiment results to original data set. We will provide analysis results and interpretation of the results with further research directions.**

***Index Terms*—Community of game users, data mining, online reviews, STEAM, usefulness of reviews.**

I. Introduction

Recently, development of internet environment and emergence of new media makes it easier for people to access to online reviews and leads to a massive amount of reviews in various forms. Thus, consumers increasingly tend to seek for peer reviews or recommendations from experts rather than relying on unilateral product information provided by companies when purchasing products. This is because consumers want to be confident in their purchasing decisions and therefore they try to reduce their perceived risk by searching for a variety of information as much as possible [1]. According to a survey conducted by Korea Internet & Security Agency in 2010 asking consumers whether they are influenced by peer reviews when making purchasing decisions, 74% of the respondents reported that peer reviews had an impact on their decision making. As described above, online reviews have a significant impact on businesses creating their own images and play a pivotal role in decision making by consumers [1]. Much research has been made on online reviews of various areas to catch up with such trends whereas industries related to 'games' which are representatives of experiential products suffer from lack of such research. It's because game communities created by users have a broad range of data in considerably diverse forms and existing tools to save and manage such data or techniques to analyze them are not enough to capture their scope and variety [2]. Recent trends show that a large volume of data is created at a high velocity and technology has developed to analyze and process a variety of big data that cannot be compared to existing ones in their diversity. Thus, people began to have interest in what have been abandoned without proper management and try to find out business values in those data. In practice, as data mining techniques develop to analyze a variety of both structured and unstructured forms of data, many cases have been found in Korea and other countries with which businesses use such data mining techniques to identify patterns or knowledge that have not been known so far and take advantage of them in their decision making [3].

The study was intended to apply various data mining techniques to community data from games available on STEAM and analyze factors that have an impact on reviews uploaded by users on their communities in order to understand community environment in which users are engaged in and suggest practical measures to enhance the usability of such communities and loyalty of users.

This paper is organized as follows: 1) Section II covers related works of this research. 2) We propose a research framework of the study in Session III. 3) Session IV addresses a case study in accordance with the framework are suggested in Session III and, finally, 4) We summarize the research results in Session V. We also provide some research limitations and future research directions with practical implications.

II. Related Works

A. Online Review

Consumers typically search for a variety of information to be confident in their decision when purchasing products. They actively search for information such as user experiences and reviews provided by other consumers [4], and it is called online word-of-mouth [5]. Bickart (2001) defines word-of-mouth as 'online information exchange among consumers regarding product information, user experiences and recommendations' [6]. Online reviews are representing

Manuscript received October 12, 2016; revised December 29, 2016. This research was supported by the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the CPRC(Communication Policy Research Center) support program (IITP-2016-R08801610080001002) supervised by the IITP(Institute for Information & communications Technology Promotion)"

Ha-Na Kang and Hye-Ryeon Yong are with Graduate School of Interaction Design, Hallym University, Chun-Cheon, Republic of Korea (e-mail: khnnn0607@naver.com, yong-@naver.com).

Hyun-Seok Hwang is with the Business Administration Department, Hallym University, Chun-Cheon, Republic of Korea (e-mail: hshwang@hallym.ac.kr).

doi: 10.18178/ijimt.2017.8.2.709

forms of online word-of-mouth which refer to reviews of products or services posted online by customers. Online reviews are considered to have relatively higher credibility compared to unilateral information provided by companies as people tend to have trust in information provided by consumers who have used products rather than commercial sources by companies [1]. Online reviews can be found for most of the products available in online markets including books, clothes, home appliances, furniture, games and music and such reviews are perceived to be an important factor that has an impact on purchasing decision by consumers and images created by businesses [7]. According to Forrest Research, an online market research company, more than half of the customers visiting online shopping malls reported that they refer to online reviews posted by other consumers [8]. As described above, online reviews play an important role in purchasing decision and therefore academic approach is required to investigate strategic values for practical use of online reviews. Thus, the study analyzed factors that have an impact on helpfulness of game reviews based on data provided by STEAM game community.

B. Usefulness of Reviews

As internet environment develops and a massive amount of review information is created at a rapid rate, characteristics of online consumer reviews are getting more attentions as they can offer increased helpfulness to consumers. Since 1995, Amazon.com, Inc. has introduced a review and evaluation system to encourage its users to participate in it. The system helps users to distinguish useful reviews from the ones that are not useful to them. Research on helpfulness of online reviews analyzes factors and patterns that have an impact on it to identify helpfulness of reviews [9]. Mudambi (2010) analyzed online reviews on Amazon.com to examine the effect of review ratings and the number of words in reviews on their helpfulness [10] and Kwon Jin-young (2016) also used review data on Amazon.com and classified them into structural characteristics and reviewer characteristics to analyze factors influencing helpfulness of reviews. The study investigated STEAM game community in which game users upload their reviews and other users can make their votes whether the reviews are 'helpful or not' just like they do on Amazon.com. Thus, the study was intended to analyze helpfulness of reviews based on data from STEAM game community.

C. STEAM

STEAM is a gaming platform developed by Valve Corporation offering services related to digital distribution, digital rights management (DRM) and multiplayer gaming. It made an official release on September 12, 2003 and has grown to be the world's largest gaming platform. STEAM covers a variety of game genres including first-person shooters (FPS), role playing, racing and even independent games for their digital management and distribution. Payments are available in diverse currencies including USD, EUR, GBP, YEN and KRW for thousands of game titles and once payments are made, licenses are registered to user library. Users can easily download and enjoy games from their library with their account information in anywhere at any time and games registered on STEAM are automatically updated in real time.

III. Research Framework

In order to analyze helpfulness of user reviews in STEAM game community, we suggested the following framework: i) In the first step, game data is selected and gathered and then ii) outliers and missing values are removed and corrected for preprocessing in next step. iii) Data set is randomly split into 10 mutually exclusive subgroups iv) Next step is adopting data mining techniques. iv) We compare prediction accuracies of results using 10-fols cross validation method. The study applied CART, Neural Network and Multi-class SVM as data mining techniques to analysis as they can be used to predict dependent variables. The procedure of the study is illustrated in Fig. 1.

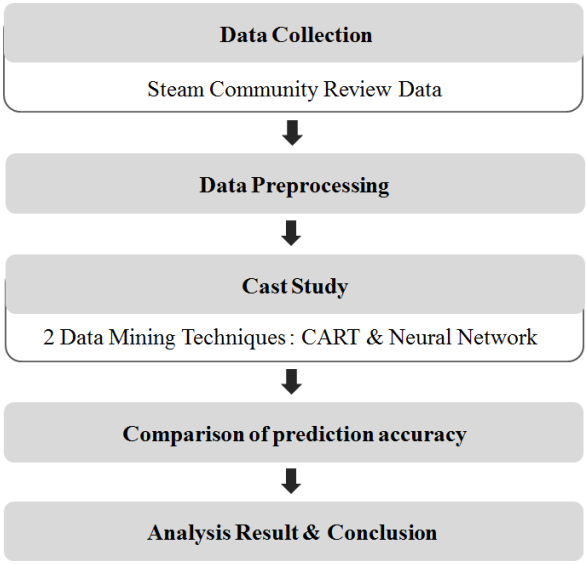

Fig. 1. Research Framework.

IV. Case Study

A case study is conducted based on the framework suggested in Chapter 3. First, community data were selected from games available on STEAM and then characteristics of variables were analyzed.

A. Data Description

Based on the data used in the study, it was intended to analyze factors that have an impact on helpfulness of user reviews in game communities. The game data used in the study were based on open source data available on GitHub (https://github.com/mulhod/STEAM_reviews) and the data were collected from STEAM, an online gaming platform. Since STEAM is the largest digital gaming platform in the world, it is proper to use review data on STEAM for the purpose of analyzing helpfulness of user reviews on games.

Research variables and their definitions are shown in Table 1. In the study, 79,437 game review data were collected from 11 different games available on STEAM and num_found_helpful_percentage was selected as a dependent variable. This variable refers to the ratio of votes made to 'helpful' when users make their votes on whether specific reviews were 'helpful or not.'

TABLE I: VARIABLE DESCRIPTIONS

VARIABLES	OPERATIONAL DEFINITIONS	ROLE
ID_NEW	UNIQUE NUMBER FOR EACH USER	IDENTIFIER
USER_NAME	USER'S ID	NONE
NUM_FOUND_HELPFUL_PERCENTAGE	RATIO OF VOTES MADE TO 'HELPFUL'	TARGET
NUM_FRIENDS	NUMBER OF FRIENDS IN THE USER'S STEAM	INPUT
NUM_REVIEWS	USER-UPLOADED REVIEWS	INPUT
NUM_SCREENSHOTS	NUMBER OF GAME SCREENSHOTS UPLOADED BY THE USER	INPUT
NUM_VOTED_HELPFULNESS	TOTAL VOTES	INPUT
NUM_WORKSHOP_ITEMS	NUMBER OF USER'S WORKSHOP ITEMS	NONE
RATING	USER'S RECOMMENDATION FOR THE GAME (NOT OR RECOMMEND)	INPUT
TOTAL_GAME_HOURS_LAST_TWO_WEEKS	USER'S TOTAL GAME TIME OVER THE LAST TWO WEEKS	INPUT
POS	POSITIVE ODDS FOR REVIEW	NONE
NEG	NEGATIVE ODDS FOR REVIEWS	NONE
GAME_NAME	GAME NAME	NONE
GENRE	GAME GENRE	NONE
GROUP	RANDOMIZED GROUP	NONE

※ num_found_helpful_percentage : Dependent variable

B. Data Preprocessing

79,437 review data in text format collected for the study were preprocessed so that they can be used for analysis. After merging 11 game review, we remove outliers and missing values. As a result, a total of 41,164 data with 13 variables were used in the study. They were randomly divided into 10 mutually exclusive subgroups for 10-fold cross validation method.

C. Data Mining Techniques

IBM SPSS Modeler was used to analyze factors influencing helpfulness of reviews. Two different data mining techniques were used for analysis and the prediction accuracies of the techniques are compared. Mean absolute error (MAE) and Sum of Square Error (SSE) were used to evaluate the prediction performance of two techniques.

1) CART

CART is a kind of decision tree algorithms. Decision Tree uses a tree-like model that predicts the value of a target variable based on several input variables. Decision Tree breaks down a dataset into smaller subsets that contain instances with homogeneous values.

CART searches multiple candidate split criteria and choose the best one maximizing the purity of split nodes. The branches grow until a certain branch has only instances with a unique category. Thus, in order to minimize the probability of overfitting, branches in over-grown trees should be properly pruned to maintain an appropriate tree size.

In the study, we set minimum percentages of instances in a branch as a stopping rule to prevent an over-grown tree. The minimum number of instances for parent nodes was 2% of whole sample size while the minimum number of child nodes was 1%. Another stopping rule is the maximum height of the tree and it is set to 5.

Fig. 2 illustrates the analysis result created by CART and num-voted_helpfulness is identified as a variable that has the most significant impact on the dependent variable, num_found_helpful_percentage.

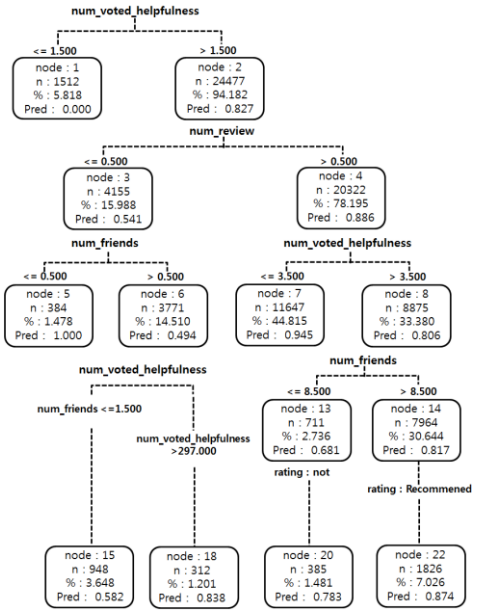

Fig. 2. Result of CART.

2) Artificial neural network

Artificial Neural Networks are processing algorithms that are simply modeled after the neuronal structure of human brain. Artificial Neural Network is a computing system made up of highly interconnected processing artificial neuron in multiple layers, which process inputs based on their dynamic interconnect weights to gain output(s).

It is also a data mining technique, just like the decision tree, in which collected data are subject to repeated learning process to find out new patterns embedded in them [12]. The neural network does not rely on complex algorithms but adjusts weighted values that connect neurons within the network in a non-algorithmic and unstructured manner to reach a solution [13].

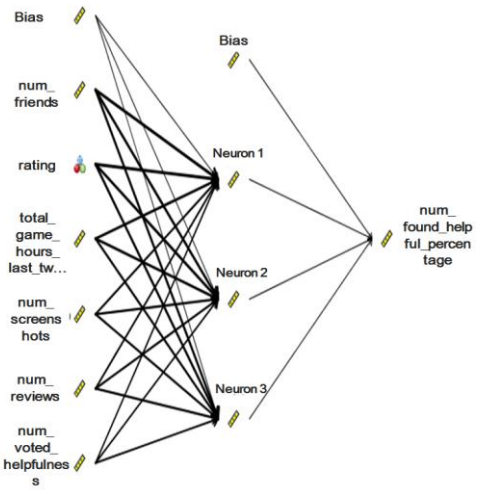

Fig. 3. Result of neural network.

The neural network has two important benefits in that the data are not assumed to be distributed normality [14]. The

neural network model was set to a multilayer perceptron (MLP) and the stopping rule was set to 15 minutes so that maximum training time can be used. ANN in this research has one input layer, one output layer and one hidden layer with 4 nodes.

Fig. 3 shows the result of analysis by the neural network and num_friends was the variable with the biggest impact on the dependent variable.

V. Research Result

A. Comparison of Prediction Accuracy among Analytic Techniques

To overcome the limited number of cases in the mashup dataset, 10-fold cross validation are used in comparing the accuracy of multiple mining techniques. One round of cross-validation includes partitioning a stratified sample of data into complementary subsets, performing the analysis on one subset (called the training set consists of 90% of the sample), and validating the analysis on the other 10% subset (called the test set). To enlarge sample size and to minimize variability, 10 rounds of cross-validation are performed using mutually exclusive partitions.

Then, MAE and SSE which can evaluate the prediction performance of the models were calculated to identify the degree of errors by the two models. MAE indicates the average of the errors while SSE considers the deviations of the errors. The formula to indicate prediction errors is shown in Fig. 4.

$$\text{MAE} = \frac{1}{n}\sum abs(V_A - V_P)$$

$$\text{SSE} = \sum abs(V_A - V_P)^2$$

※ V_A = Actual Value, V_P = Predictive Value

Fig. 4. Formula of error measures.

TABLE. II: Prediction Errors

Method	MAE	SSE
CART	0.112E+6	1213.269
Neural Network	0.188E+3	2691.165

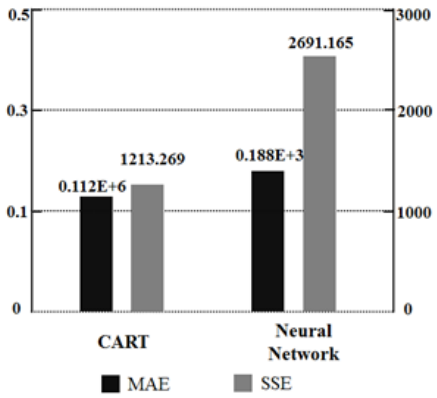

Fig. 5. Prediction error.

The result of comparison for prediction errors between CART and Neural Network is described in Table. II and Fig. 5.

B. Excellent Prediction Model and the Important of Variables

In the study, CART was used to analyze factors that have an impact on helpfulness of game reviews as it showed the highest prediction accuracy. The importance of variables drawn from CART is illustrated in Fig. 6.

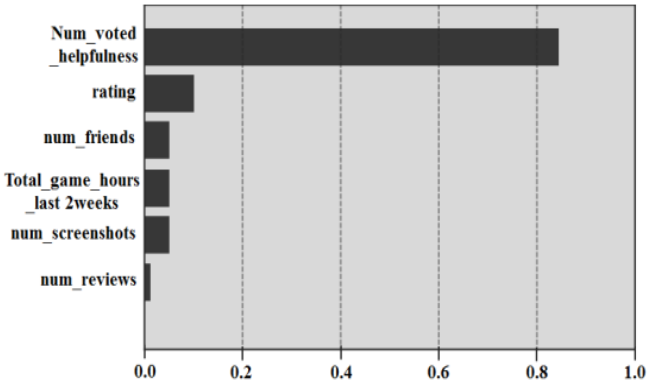

Fig. 6. Importance of variables in CART.

C. Analysis Result

The result of analysis by CART which showed the highest prediction accuracy suggested that the total number of votes, num voted helpfulness, was the variable with the most significant impact on helpfulness of reviews followed by the rating, num_friends indicating the number of friends on the community and total_game_hours_last_two_weeks showing the game playing hours by reviewers. The rating was a variable that indicates whether users recommend games that they played or not and it has an impact on helpfulness of reviews on the condition that users recommend what they played. The number of screen shots (num_screenshots) and the number of reviews (num_reviews) also had positive correlations with helpfulness of reviews.

VI. Conclusion

The study aimed to identify factors affecting helpfulness of reviews uploaded by users on the communities by analyzing unrefined game data with data mining techniques. Two data mining techniques (CART and Neural Network) were used to analyze the game review data set and CART was identified as a better method to predict the target variable, num_voted_helpfulness. The most significant variable is num_found_helpful_percentage, followed by the rating, num-friends, total_game_hours_last_two_weeks, num_screenshots and num_reviews.

In the study, research variables are limited to characteristics of reviewers such as the number of friends, the number of votes and total game hours were analyzed. However, future research will focus on structural characteristics of reviews such as emotional words, explanative words, the length of reviews and the date of reviews registered to identify their impact on helpfulness of reviews. In addition, only two prediction models, CART and Neural Network, were used in the study but future research will expand to include other types of prediction modes such as SVM and random forests for comparison and verification.

REFERENCES

[1] H. G. Lee and H. Kwak, "Investigation of factors affection the effects of online consumer reviews," *Journal of Information Policy*, vol. 20, no. 3, pp. 3-17, 2013.

[2] H. R. Yong, D. J. Kim, and H. S. Hwang, "A study of analyzing realtime strategy game data using data mining," *Journal of Korea Game Society*, vol. 15, no. 4, pp. 59-68, Aug 2015.

[3] J. G. Kim, "Domestic & foreign big data trends & success stories," *Industrial Engineering Magazine*, vol. 23, no. 1, pp. 47-52, Mars 2016.

[4] R. A. Peterson and C. M. Maria, *Consumer Behavior*, 7th edition, Upper Saddle River, NJ: Prentice-Hall, 2003.

[5] N. Thopson, *More Companies Pay Heed to Their 'Word of Mouse' Reputation*, New York Times, June 2003.

[6] B. Bickart and R. Schindler, "Internet forums as influential sources of consumer information," *Journal of Interactive Marketing*, vol. 15, no. 3, pp. 31-40, 2001.

[7] J. H. Kim, H. S. Byeon, and S. H. Lee, "Enhancement of user understanding and service value using online reviews," *Journal of Korea Information Systems Society*, vol. 20, no. 2, pp. 21-36, 2011

[8] Y. Chen and J. Xie, "Online consumer review: Word of mouth as a new element or marketing communication mix," *Management Science*, vol. 54, no. 3, pp. 477-491, 2008.

[9] B. Liu, *Sentiment Analysis and Opinion Mining (Synthesis Lectures on Human Language Technologies)*, Morgan & ClayPool, 2012.

[10] S. M. Mudambi and D. Schuff, "What makes a helpful online review? A study of customer reviews on Amazon.com," *MIS Quarterly*, vol. 34, no. 1, pp.185-200, 2010.

[11] J. Y. Kwon and M. Y. Lee, "A study on the determining factors of online review helpfulness," *Journal of Korea Intelligent information Systems Society*, pp. 205-211, 2012.

[12] V. B. Rao and H. V. Rao, "C++ neural network and fuzzy logic," *New York: Management Information*, p. 408, 1993.

[13] T. Hengl, "Neural network: A neural computing primer," *Personal Computing Artificial Intelligence*, vol. 16, no. 3, pp. 32-43, 1993.

[14] J. R. Jensen, F. Qiu, and M. Ji, "Predictive modeling of coniferous forest age using statistical and artificial neural network approaches applied to remote sensing data," *International Journal of Remote Sensing*, vol. 20, no. 14, pp. 2805-2822, 1999.

Ha-Na Kang is a master's candidate in Interaction Design School at Hallym University. She majored in business administration at Hallym University. She is a research associate in CPRC (Communication Policy Research Center). Her research interests cover social network analysis, interaction design and big data analytics.

Hye-Ryeon Yong is a master's candidate in Interaction Design School at Hallym University. She majored in business administration at Hallym University. She is a research associate in CPRC (Communication Policy Research Center). Her research Interests include data mining, interaction design and big data analytics.

Hyun-Seok Hwang is a professor of business administration and a research fellow of Hallym Business Research Institute at Hallym University, Chuncheon, South Korea. He received his PhD in industrial engineering from the Pohang University of Science and Technology (POSTECH), South Korea. His current research focuses on big data analytics, opinion mining, interaction design.