Text Mining and Sentiment Analysis for Predicting Box Office Success

Yoosin Kim[1], **Mingon Kang**[2], **Seung Ryul Jeong**[3],
[1] Big Data Analytics Department, University of Seoul, Seoul, Korea
[2] Department of Computer Science, Kennesaw State University, GA, U.S.A.
[3] Business IT Graduate School, Kookmin University, Seoul, South Korea
*Corresponding author: Seung Ryul Jeong

Received March 5, 2018; accepted April 12, 2018; published August 31, 2018

Abstract

After emerging online communications, text mining and sentiment analysis has been frequently applied into analyzing electronic word-of-mouth. This study aims to develop a domain-specific lexicon of sentiment analysis to predict box office success in Korea film market and validate the feasibility of the lexicon. Natural language processing, a machine learning algorithm, and a lexicon-based sentiment classification method are employed. To create a movie domain sentiment lexicon, 233,631 reviews of 147 movies with popularity ratings is collected by a XML crawling package in R program. We accomplished 81.69% accuracy in sentiment classification by the Korean sentiment dictionary including 706 negative words and 617 positive words. The result showed a stronger positive relationship with box office success and consumers' sentiment as well as a significant positive effect in the linear regression for the predicting model. In addition, it reveals emotion in the user-generated content can be a more accurate clue to predict business success.

Keywords: Text Mining; Sentiment Analysis; Prediction; Box office Success; Word of Mouth

A preliminary version of this paper was presented at ICONI 2017, and was selected as an outstanding paper.

http://doi.org/10.3837/tiis.2018.08.030 **ISSN : 1976-7277**

1. Introduction

After the emergence of online communication, unstructured text data such as customer reviews, forum messages, blogs, and tweets, has directly increased as big data and has also strongly influenced many kinds of businesses. Text data written by user is called electronic word-of-mouth or user-generated content. The unstructured text data with the uncountable volume involves rich information to understand the users' behaviors, emotions, interests, complaints, and opinions. Hence, a number of researchers have recently attempted to mine the user generated digital content such as customer reviews, blogs, and social media content, thereby emotion mining, opinion mining including sentiment analysis, has been frequently adapted in various business fields [1], [2]. Moreover, text mining and sentiment analysis, has been applied into various experiments for businesses such as box-office, music album sales, hotel reservation, restaurant popularity, and even the stock market. In particular, the film market is one of the industries for which word-of-mouth effectively works to increase sales revenue since the willingness of potential customers is affected by the reputation of people who have already watched the movie [3].

Many researchers in prior studies have conducted the lexicon-based approach and machine learning method based approach [4]–[7]. The machine learning method approach uses classifying algorithms for sentiment classification. The lexicon-based approach is to determine the polarity of the content through a linguistic resource called a sentiment lexicon such as general sentiment dictionaries(e.g., SentiWordNet and OpinionFinder) to determine the polarity of the content [3], [8] and an object-oriented sentiment dictionary. A few studies applied the domain-specific lexicon and insisted that it could lead to higher performance than general dictionaries [9]. However the studies have inspired us to develop a sentiment lexicon for better performance in emotion mining, there are rarely references to tools, techniques, and methods to build a domain-specific sentiment dictionary. Besides, Korean word-of-mouth data has seldom been examined in sentiment analysis because Korean alphabet "Hangeul" is an agglutinative language, which makes it difficult to analyze morphemes within corpora [10].

Therefore, we tried to investigate the usability of emotion mining for business decision making, and targeted the Korean film market, box office success of movies is seriously affected by electronic word-of-mouth. For the emotion mining with Korean text content, we first collected 429,193 of consumer reviews in a movie rating website by a XML crawler in R software, and it was qualified as 226,264 reviews of 144 movies. We thus conducted emotion mining such as developing a Korean sentiment dictionary for movie domain, evaluating the lexicon, and validating a prediction model for box office success. In experiments, we applied feature extraction, term definition, and polarity assignment to develop a movie sentiment dictionary. In addition, we conducted sentiment analysis using the generated dictionary and employed a linear regression model in empirical test. We expect our analysis and findings can show the usability of emotion mining of consumer generated content for prediction of business success, especially marketing sales side as well as the feasibility of the domain-specific lexicon.

2. RELATED WORKS

After the emergence of web 2.0 technologies, online communication and social networking became a user friendly means for people to connect with each other. A huge amount of

online content has been generated from numerous websites, forums, blogs, communities, and social networking sites. Especially, the mobile technologies enable people to connect and communicate anywhere and anytime, and the numberless digital content has been produced continuously. User-generated content and electronic word-of-mouth is considered to play an important role in the sales and marketing performance of real businesses, therefore organizations have exploited online media to understand consumers. Furthermore, researchers have attempted to mine authors' interests, sentiments, complaints, and opinions from text data such as customer reviews and rankings, forum messages, blogs, tweets, and news articles in online channels. It includes e-commerce sites, online forums, the blogosphere, and social networking sites within various industries doing marketing sales activities with movies, books, music albums, hotels, and restaurants.

The film market, in particular, is one of the industries for which word-of-mouth effectively works to increase sales revenue since the willingness of potential customers is affected by the reputation of people who have already watched the movie [3]. Liu et al. (2010) analyzed the online contents of 257 movies in Yahoo Movies using the sentiment dictionary, Pang et al. (2002b) tried to judge the sentiment of movie reviews applying several machine learning algorithms, and Rui et al. (2013) revealed the effect of tweets on movie sales using sentiment analysis. As the previous research, the consumer has been influenced by online reputation [11].

By the way, sentiment analysis is often used in opinion mining which focuses on extracting authors' opinions and sentiments from user-generated content such as customer reviews, forum messages, and blogs [1], [12], [13]. Methods for extracting sentiment and opinion from text can be divided into the following two approaches: the lexicon-based approach and the machine learning approach [10]. The machine learning approach generates classifiers by training data set with labels and the lexicon-based approach applies a linguistic resource called a sentiment dictionary or sentiment lexicon.

In the lexicon-based sentiment analysis, general sentiment dictionaries such as SentiWordNet, SenticNet and OpinionFinder, have been frequently used for sentiment classification because there is the advantage that the dictionary is easy to obtain and is more accurate in general domains [14]. A case study using SentiWordNet evaluated the business constituents of Wal-Mart [15] and generated opinion analysis results concerning traffic dynamics, topic evolution, the sentiment of active topics, and the opinion leader analysis. Liu et al. provide a word-of-mouth analysis of managerial decisions in the Hollywood film market, employing SentiWordNet and OpinionFinder to extract valence and numbers of words (2010). Bollen et al. collected over nine million tweets over 10 months from February 28 to December 19, 2000 and determined the mood of each tweet, tagged by OpinionFinder and Google-Profile of Mood States [16].

However, some researcher insisted an object-oriented sentiment dictionary called a domain-specific lexicon can lead to higher performance than general sentiment dictionaries such as SentiWordNet and OpinionFinder [9]. Rao et al. (2013) proposed an algorithm to generate a word-level emotional dictionary for social emotion extraction and a topic-level dictionary for relationships between topics and social emotions. They concluded that a dictionary generated for certain purposes can be more efficient for predicting the emotional distribution of news articles. Another research studied an approach to automatic emotion recognition through semantic labeling in an emotional text corpus by the a priori algorithm[17]. Despite of the research, however, general dictionaries have been preferred than manual sentiment

dictionaries because of their high coverage and reliability [18]. Besides, Korean word-of-mouth data has seldom been examined in sentiment analysis because Korean is an agglutinative language, which makes it difficult to analyze morphemes within corpora [10].

Therefore, we attempted to examine the usability of emotion mining for business decision making. As the target domain, we set the Korean film market, box office success of movies is seriously affected by electronic word-of-mouth. For the emotion mining with Korean text content, first we collected 429,193 of consumer reviews in a movie rating website by a XML crawler in R software, and it was qualified as 226,264 reviews of 144 movies. And we thus conducted experiment such as developing a Korean sentiment dictionary for movie domain, evaluating the lexicon, and validating a prediction model for box office success. We applied feature extraction, term definition, and polarity assignment to develop a movie sentiment dictionary. In addition, we conducted sentiment analysis using the generated dictionary and employed a linear regression model in empirical test. We expect our analysis and findings can show the usability of emotion mining of consumer generated content for prediction of business success, especially marketing sales side as well as the feasibility of the domain-specific lexicon.

3. APPROACH

For examining the usability of emotion mining for business decision making, we develop a movie domain-specific sentiment lexicon with the electronic word-of-mouth data written in the Korean alphabet “Hangeul”, and test the performance of the lexicon in predicting box office success. First, to make a movie sentiment dictionary, 1) gather the electronic word-of-mouth data in a famous movie rating website, and extract linguistic terms from the gathered data through natural language processing; 2) select preliminary terms in a proper set size via sparsity and SVM; 3) the polarity of the terms were determined by probability; 4) the preliminary dictionary was tested in a lexicon-based sentiment classification method.

Next, to predict box office success with the lexicon, we conduct the empirical test to validate correlation and influence between sentiment analysis results and movie sales performance, which includes the viewer numbers, the number of released screens, and total revenue. Electronic word-of-mouth data was collected from a famous movie rating website in Korea (http://movie.daum.net). The data included movie reviews and popularity scores rated by online users. In this website, online users can write reviews and rate the popularity score of the movies. A review is allowed up to 140 letters and evaluated with a rating score between zero (0), meaning extremely negative, and ten (10), the highest positive score.

For gathering the electronic word-of-mouth data, we developed a crawling software program using an XML package in R and gathered a total number of 429,193 reviews from December 2013 to January 2015. The data mentioned over 15,000 movies, even including old movies and home videos. Therefore, we cleaned up the data for the next step. Some movies, which were released prior to the collection date, were excluded, and movies having fewer reviews under 500 were also removed from the data set. Finally, the word-of-mouth data was qualified as 233,631 reviews of 147 movies. **Table 1** describes the number of movie reviews across the popularity scores from the worst, zero, to the best, ten, and shows that about 70% reviews were rated as eight to ten.

Table 1. Collected movie reviews

Popularity	0	1	2	3	4	5	6	7	8	9	10
Number	12,489	9,486	4,405	5,543	5,037	10,402	9,754	13,800	21,375	30,782	110,558
Rate (%)	5.35	4.06	1.89	2.37	2.16	4.45	4.17	5.91	9.15	13.18	47.32

We also investigated market sales data of those movies from the Korean Film Council (http://www.kobis.or.kr), which is an organization, entrusted from the Ministry of Culture, Sports and Tourism of the Republic of Korea. The data includes the total revenue, the number of screens, the viewer numbers, the release date, and so on.

4. KOREAN SENTIMENT DICTIONARY IN MOVIE DOMAIN

We defined four steps to make a Korean sentiment dictionary for the films. The procedure include extracting linguistic terms from the gathered data, selecting preliminary terms in a proper set size, determining the polarity of terms, and testing the preliminary dictionary in a lexicon-based sentiment classification.

4.1 Extracting Terms

The first step to build the movie specific lexicon is extracting candidate terms that might have sentiment polarity, such as “pros and cons” or “positive and negative.” For the extraction, we made a sample data set by combining extremely negative reviews rated at a popularity score of zero and positive reviews ranked at ten, the best score, by online users. The sample data set included 10,000 reviews, merging 5,000 reviews randomly sampled from each polarity group. After the sample data was selected, we performed natural language processing on the review content. We eliminated unavailable letters (e.g., kkk and hhh), and characters such as emoticons (e.g., -.-, ^^, ☹), numbers, and punctuation. Then, we conducted a decomposition process, in which sentences were parsed at the term level, serving to eliminate useless data. Consequently, 14,418 terms of 9,179 reviews remained in the sample data set. All terms could be used in the lexicon; however, they had to be selected for the purposes of analysis efficiency. In this processing for term extraction, we used “KoNLP” (the package for Korean natural language processing) and “tm” (the package for text mining) in the R project.

4.2 Selecting Terms and Set Size

Our goal in this step is to select a preliminary lexicon in the proper set size. We attempted to define a proper set size and select a preliminary lexicon from the 14,418 terms extracted in the first step, to maximize sentiment classification performance. In other words, we tried to find how many sentiment terms should be included in the sentiment dictionary. The previous research achieved 81.4% accuracy in sentiment analysis using 2,633 features as top frequency words and the accuracy was close to the best performance 82.9% using 16,165 terms [7]. Therefore, following this research, we selected more frequent terms in the data using a function of NLP which removes sparse terms, and conducted SVM experiments to fine the best performance set size. That selected feature sets were 4 groups from a set with 405 terms to a group with 1,542 terms. In this experiment, we did the 10-fold cross-validation on each data set and the result showed that 1,542 terms achieved the best performance (81.12%) in sentiment classification of movie reviews (see **Table 2**).

Table 2. Experiments by terms size

	Sparsity	No. of terms	Accuracy (10-fold cross-validation)		
			Min	Max	Average
Ex1	.9980	405	75.16	81.13	77.93
Ex2	.9990	776	77.56	81.26	79.77
Ex3	.9992	1107	77.67	81.79	80.40
Ex4	**.9995**	**1542**	**79.19**	**82.35**	**81.12**

4.3 Determining the Polarity of Terms

The next step was determining the polarity of terms, which are selected in the previous phase. For determining polarity of each term, we applied the probability where the word existed more frequently. We calculated the frequency of the word in both negative and positive reviews, and assigned the word into the dominant sentiment by the follow formula:

IF Word(i)apprFreq_Pos > Word(i)apprFreq_Neg THEN
Word(i)senti = Positive;
ELSEIF Word(i)apprFreq_Pos < Word(i)apprFreq_Neg THEN
Word(i)senti = Negative;
ELESIF Word(i)apprFreq_Pos = Word(i)apprFreq_Neg THEN
Word(i) = NULL;
END IF;

Through the sentiment assignment, we defined 677 positive sentiment words, 784 negative sentiment words, and 81 words with the same number of appearances on both sides. Then, we excluded neutral words, actors' names, and duplicated terms, which yielded a preliminary lexicon of 1,313 terms consisting of 706 negative words and 617 positive words. **Table 3** shows a few sample words of the movie domain sentiment lexicon. Terms expressing positive emotions included "happy", "fun", and "impressive". On the other hand, negative words such as "trash", "waste", and "devil", clearly disclosed the authors' unpleasant feelings towards the movies.

Table 3. Sample of movie sentiment lexicon

Positive Words	Negative Words
Enjoyed (잘봤다), Fun (재밌다), Well made (잘만든), Impressive (감동적), Worthy (볼만한), Awesome (짱짱), Recommend (추천)	Trash (쓰레기), Waste (돈 아까운), Devil (도대체), Cheapie (싸구려), Tricked (낚이다), sham (엉터리), Disappointed (실망)

4.4 Testing the Preliminary Dictionary

After polarity assignment, we regarded the classified terms as a preliminary sentiment dictionary to apply into sentiment analysis of movie reviews. We tested the preliminary sentiment dictionary through sentiment lexicon-based classification of the sample data set. The sentiment of a movie review was evaluated by the distances between the number of appearances of the positive and negative terms. We proposed that if the summation of appearances of positive terms was greater than that of negative terms, then the review was classified as having a positive sentiment, etc. Then, we converted each polarity to a

numerical value for calculation, i.e., the positive sentiments had the value one (1) and the negative sentiments zero (0).

IF Sum(appr_Pos_term) > Sum(appr_Neg_term) THEN
Word(i)senti = Postive(1);
ELSE Sum(appr_Pos_term) = < (appr_Neg_term) THEN
Word(i)senti =Negative(0);
END IF;

To assess the classification performance, we measured recall, precision, accuracy α, and F1-scores [10], [19], [20]. Total accuracy α is defined as the percentage of sentiments correctly predicted of the total instances. F1 score is the harmonic mean of macro-averaged precision and recall. Precision is the proportion of correct instances of the total instances predicted, and recall is the proportion of instances correctly predicted of the real instances. A higher result means better performance, but the two measurements generally have an inverse relationship. The sentiment classification result was compared with the popularity score, i.e., Pros (10) or Cons (0), of each review, as seen in **Table 4**. The preliminary lexicon achieved 81.69% classification accuracy and 88.08% of F1 score.

Table 4. Result of the preliminary lexicon-based sentiment analysis

Predict \ Act	Pros(10)	Cons(0)	n	Accuracy	F-1
Positive	3,413	502	9,179	0.8139	0.8808
Negative	1,206	4,058			

Comparing to the previous research result, which showed 81.4% accuracy using 2,633 features [7], the performance of the dictionary in sentiment analysis seemed to be successful, although the preliminary lexicon was smaller, containing 1,313 sentiment terms.

5. BOX-OFFICE SUCCESS PREDICTION

In this section, we define the statistical measurement to predict box office success with sentiment analysis result with the dictionary and conduct the empirical test.

5.1 Measurement for Box-office Success Prediction

According to the previous studies, a movie's success and its word-of-mouth transmission have a strong positive relationship [3], [5]. In this regard, we considered that the sentiment analysis result of a movie's reviews can show a movie's box office success. For the empirical validation, we conducted sentiment analysis with the generated sentiment dictionary, and investigated the correlation and influence of the electronic word-of-mouth to the market performance such as the viewer numbers, the number of theaters, the revenue, etc. To analyze these relations, we defined variables for a multiple linear regression model (see **Table 5**). There are general variables such as total customers, total revenue, and total screens as well as the newly defined variables such as csmScrn, womAmt, womMean, and sntMean. The csmScrn is the number of viewers divided by the number of theaters; the womAmt is the amount of reviews for a movie; the womMean is the average of the popularity scores, as rated by online users about the movie; the sntMean is the average of the result of the sentiment analysis of review contents generated by users.

Table 5. Key variables used in sentiment analysis validation

Total customers	Total number of customers(audience) of a movie
Total Revenue	Total revenue (translating from Korean Won to U.S. Dollar) of a movie
Total Screens	Total number of Screens released a movie
csmScrn	Average number of viewers per screen to fairly calculate the popularity of each movie, disregarding budget and marketing input
womAmt	Total number of WOM mentioning a movie and rating the popularity of the movie
womMean	Mean of the popularity score of each movie
sntMean	Mean of the score analyzing WOM sentiment of each movie by the generated lexicon

More specifically, the "csmScrn" is developed a new variable regarding as box office success because we considered box office success is strongly related with the occupancy rate of seat. Typically, low-budget films shared fewer release screenings, total revenue, and customers than big-budget movies, but we cannot conclude that they are poorer than blockbusters despite of lower performance in them. For example, as shown below in **Table 6**, "Transformers: Age of Extinction," for the production of which USD 210 million was invested , was released on 1,602 screens and achieved 5.3 million viewers and USD 44 million revenue in the Korean film market. On the other hand, "Begin Again" with just an USD 8 million budget, received USD 2.7 million revenue and 3.4 million viewers on 213 screens in the Korean market. Worldwide, it grossed over USD 63 million, and in addition, the movie received mostly positive reviews from critics and audiences. Even though the big-budget movie, Transformers, had more revenue and customers than the low-budget movie, Begin Again, there is no doubt as to which one was more successful.

Table 6. Comparison of two movies

	Transformers	Begin Again
Release Date	May 25, 2014	Aug 13, 2014
Production Budget	USD 210 million	USD 8 million
Total Revenue	USD 44 million	USD 27 million
Total Customers	5,295,929	3,427,739
Total Screens	1,602	213
Total Reviews	1,733	1,130
Popularity Mean	6.40	8.53
Sentiment Mean	2.83	6.56

Source: Korean Film Council (www.kofic.or.kr)

5.2 Sentiment Analysis and Statistics

Using the generated sentiment dictionary, we conducted sentiment analysis and statistics with the movie review data, which amounted to 226,264 reviews of 144 movies excluding three noncommercial movies from 233,631 reviews of 147 movies. As the descriptive investigation, we conducted statistics about the variables and analyzed the correlation between them. **Table 7** shows the statistical summary of the variables, i.e., the number of customers per screening, the number of word-of-mouth reviews, the popularity score, and the

sentiment analysis result.

Table 7. Summary statistics of variables for 144 movies

Variables	Mean	Std dev	Min	Max	Note
Total customers	2,722,707	2,930,970	47,060	17,611,849	-
Total Revenue	20.7	22.2	3.3	135.7	Million $
Total Screens	665	267.3	169	1602	-
csmScrn	278	3,413.7	3,688	23,286	-
womAmt	1,571	2,543.3	517	28,827	-
womMean	7.426	1.003	4.575	9.653	0 to 10
sntMean	4.591	1.157	2.108	7.448	0 to 10

The average of the releasing screens is 665 units; the mean of the audience share, dividing total customers by total screens, is 278 customers, and the average number of movie reviews is 1,571 units. The womMean, which represents the average of the user-generated scores, is 7.426, spanning from the minimum score of 4.575 to the maximum of 9.653. On the other hand, the sntMean, with an average score of 4.591 in the sentiment analysis using the movie lexicon, is located at the center of the scope. It means that the online user tended to rate a movie's popularity generously but write the review honestly. The correlation matrix between variables shows that most variables are significantly correlated with each other (see **Table 8**). With the csmScrn as the market success variable, the womAmt is significantly correlated, at 0.439; the sntMen is next at 0.370 and the womMean follows. Meanwhile, concerning the total screen numbers, the sntMean and the womMean show negative correlations. It might be explained by the fact that the number of release screenings does not influence to consumer reputation.

Table 8. Correlation matrix

	1	2	3	4	5	6
1. Total revenue	1.000					
2. Total customers	0.997**	1.000				
3. Total screens	0.691**	0.682**	1.000			
4. csmScrn	0.787**	0.793**	0.296**	1.000		
5. womAmt	0.558**	0.572**	0.330**	**0.439****	1.000	
6. womMean	0.204**	0.210**	-0.101	**0.342****	0.201**	1.000
7. sntMean	0.164**	0.181**	-0.210*	**0.370****	0.173**	0.662**

**and* denote significance levels at 1 % and 5%, respectively

We estimated the linear regression with three variables, womAmt, womMean, and sntMean, to the csmScrn. As shown in **Table 9**, the total number of reviews (womAmt) and the sentiment of the reviews (sntMean) had a positive and significant effect on box office success (csmScrn), unlike popularity score (womMean). This means that the amount of word-of-mouth and their sentiment is a more reliable factor to predict market success. In addition, the analysis on our movie data shows that a one-point increase in sentiment could bring more than 6,770 customers to a theater.

Table 9. Estimation results of linear regression analysis

Variables	Estimate	Std. Error	Pr(>|t|)
csmScrn *(dependent variable)*			
womAmt	0.51	0.10	0.00**
womMean	389.70	325.90	0.23
sntMean	6,770.00	2,809.00	0.02*
N: 144 movies , Adjusted R-squared: 0.2741, p-value: 0.00**			

**and* denote significance levels at 1 % and 5%, respectively

Consequently, we conducted a statistical analysis to validate the sentiment lexicon with market performance. The mean of the sentiment scores was 4.591 points, and it lightly leaned toward the negative area. The mean of the popularity scores rated by online-users was 7.246, and 47% of them gave 10 points. This indicates that the user-generated reviews seemed to express their straightforward feedback. For that reason, the sentiment mean has a closer relationship with box-office success and has a more significant effect than the popularity mean.

6. DISCUSSION AND IMPLICATIONS

With emerging online communications, unstructured text data such as customer reviews, forum messages, blogs, and tweets, has directly and strongly influenced most kinds of businesses. In particular, the film market is much affected by word-of-mouth if a movie is a box office hit. Thus many researchers have frequently applied emotion mining such as sentiment analysis into analyzing electronic word-of-mouth. However there is rarely research for emotion mining in box office success prediction with the Korean sentiment dictionary. In this regard, we tried to investigate the usability of emotion mining for business decision making, and targeted the Korean film market. We collected 429,193 movie-reviews written by online users in a famous movie rating website and qualified them as amounting to 226,264 reviews of 144 movies by a XML crawling program in open source R. The research applied feature extraction, term definition, and polarity assignment to develop a movie sentiment dictionary, and also conducted sentiment analysis using the generated dictionary and a linear regression model for empirical test as well. We developed the Korean sentiment dictionary including 706 negative words and 617 positive in movie domain and the dictionary achieved 81.69% accuracy and 88.08% of F1 score in sentiment classification.

After developing the sentiment dictionary, we employed the statistical analysis to validate the feasibility of the dictionary for the whole of the movie review data. We gathered the box office data and generated the audience share as the market performance variable. In this empirical test, we found several significant results of the sentiment analysis. First, the sentiment of movie reviews was exiguously close to the negative, unlike the popularity score, which leaned toward the positive. Next, in the correlation analysis, the sentiment mean had a stronger positive relation to the market success variable than the popularity mean. Lastly, the estimation of the linear regression with three variables, womAmt, womMean, and sntMean, showed that the total number of reviews and the sentiment of the reviews had a positive and significant effect on box office success, but the popularity score was insignificant in the model. As the result, it seems to conclude that the user-generated reviews are more straightforward and accurate than the popularity star ratings in movie critiques.

Consequently, this study has confirmed the feasibility that a Koran sentiment dictionary, which is purpose to predict box-office success, can ensure higher performance than even the popularity rating. In addition, the demonstrated lexicon achieved over 80% classification accuracy despite the small number of linguistic features. This result should support a more efficient and effective means of conducting sentiment analysis on big data and provide a good reference for potential users of sentiment analysis with open source packages such as the R project, as well.

Findings from this study showed that developing a Korean sentiment dictionary in a specific domain is feasible and sentiment analysis with the sentiment lexicon can be applied to predict box-office success. Researchers and business users can understand the consumers' opinions towards its own product and service through emotion mining of user-generated content. If an organization develops the online reputation monitoring system for market sensing, it can reveal consumers' feedback toward its products and services in real time. In addition, the firm could predict its market success as well as competing products or services.

7. CONCLUSION AND FUTURE RESEARCH

This study aimed to investigate the usability of emotion mining for business activities such as marketing sales. We developed the Korean sentiment dictionary in movie domain and validated the feasibility of the lexicon to predict box office success. We collected online movie-reviews in a famous movie rating website by a XML crawling program in open source R and qualified them to develop a movie specific sentiment dictionary. We employed Korean natural language processing, support vector machines algorithm, sentiment analysis, and statistics for box office success prediction. Through those processes, we extracted preliminary sentiment words, determined sentiment polarity of them, and defined the Korean movie sentiment dictionary.

Furthermore, sentiment analysis with the lexicon showed the feasibility of the movie domain sentiment dictionary with the box office success prediction model in film market. It revealed that consumer sentiment in electronic word-of mouth has a stronger positive relationship with box office success as well as a significant positive effect in the linear regression. It provides not only the theoretical reference to researchers who want to handle Korean word-of-mouth, but also practical guide to marketers and business users who want to analyze and predict box-office success.

In this regard, we believe that researchers and practitioners can consider opinion mining method for analyzing Korean content to discover consumer reputation and market insight for estimating business success. The procedure in the proposed approach describes how to collect electronic word-of mouth data, extract terms from data, select preliminary terms in a proper size set, determine sentiment polarity, and test the preliminary dictionary. Furthermore, statistical analysis for box-office success provides empirical evidence that consumer sentiment from the user-generated content predict market success in the movie industry. Business users and marketers who wish to mine consumer opinion from word-of mouth in online media can apply our approach and procedure not just within the movie industry but in other industries as well.

This study has several improvements to be addressed in future research. First, next studies will consider multi-grams and various term extraction algorithms for developing the sentiment dictionary. Next, Korean text mining and sentiment analysis can be extended to

various social media platforms such as Twitter, Facebook, LinkedIn, and Instagram. Each services has its own characteristics and users can also be distinguished through features including images, retweets, and preferences. Finally our approach should extend to other domains, such as e-commerce, high-end goods, and health care service whereby consumer reputation is regarding as a critical success factor.

Rererences

[1] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "New Avenues in Opinion Mining and Sentiment Analysis," *IEEE Intell. Syst.*, vol. 28, no. 2, pp. 15–21, 2013. Article (CrossRef Link)

[2] Y. Kim, R. Dwivedi, J. Zhang, and S. R. Jeong, "Competitive intelligence in social media Twitter: iPhone 6 vs. Galaxy S5," *Online Inf. Rev.*, vol. 40, no. 1, pp. 42–61, 2016. Article (CrossRef Link).

[3] Y. Liu, "Word of Mouth for Movies: Its Dynamics and Impact on Box Office Revenue," *J. Mark.*, vol. 70, no. 3, pp. 74–89, Jul. 2006. Article (CrossRef Link).

[4] Y. Kim, D. Y. Kwon, and S. R. Jeong, "Comparing Machine Learning Classifiers for Movie WOM Opinion Mining," *KSII Trans. Internet Inf. Syst.*, vol. 9, no. 8, pp. 3178–3190, 2015. Article (CrossRef Link).

[5] H. Rui, Y. Liu, and A. Whinston, "Whose and what chatter matters? The effect of tweets on movie sales," *Decis. Support Syst.*, vol. 55, no. 4, pp. 863–870, Nov. 2013. Article (CrossRef Link)

[6] Jin-Cheon Na, T. T. Thet, and C. S. G. Khoo, "Comparing sentiment expression in movie reviews from four online genres," *Online Inf. Rev.*, vol. 34, no. 2, pp. 317–338, 2010. Article (CrossRef Link).

[7] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proc. of Conf. Empir. Methods Nat. Lang. Process.*, pp. 79–86, 2002. Article (CrossRef Link).

[8] Y. Liu, Y. Chen, R. F. Lusch, H. Chen, D. Zimbra, and S. Zeng, "User-Generated Content on Social Media: Predicting Market Success with Online Word-on-Mouth," *IEEE Intell. Syst.*, vol. 25, no. 1, pp. 8–12, 2010. Article (CrossRef Link).

[9] Y. Rao, J. Lei, L. Wenyin, Q. Li, and M. Chen, "Building emotional dictionary for sentiment analysis of online news," *World Wide Web*, 2013. Article (CrossRef Link).

[10] J. Lim and J. Kim, "An Empirical Comparison of Machine Learning Models for Classifying Emotions in Korean Twitter," *J. Korea Multimed. Soc.*, vol. 17, no. 2, pp. 232–239, 2014. Article (CrossRef Link).

[11] G. P. Sonnier, L. McAlister, and O. J. Rutz, "A Dynamic Model of the Effect of Online Communications on Firm Sales," *Mark. Sci.*, vol. 30, no. 4, pp. 702–716, Jul. 2011. Article (CrossRef Link).

[12] B. Pang and L. Lee, *Opinion Mining and Sentiment Analysis*. 2008. Article (CrossRef Link).

[13] H. Chen and D. Zimbra, "AI and Opinion Mining," *IEEE Intell. Syst.*, vol. 25, no. 3, pp. 74–76, May 2010. Article (CrossRef Link).

[14] A. Ortigosa, J. M. Martín, and R. M. Carro, "Sentiment analysis in Facebook and its application to e-learning," *Comput. Human Behav.*, vol. 31, no. 1, pp. 527–541, 2014. Article (CrossRef Link).

[15] H. Chen, "Business and Market Intelligence 2.0, Part 2," *IEEE Intell. Syst.*, vol. 25, no. 2, pp. 2–5, Mar. 2010. Article (CrossRef Link).

[16] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *J. Comput. Sci.*, vol. 2, no. 1, pp. 1–8, Mar. 2011. Article (CrossRef Link).

[17] C. Wu, Z. Chuang, and Y. Lin, "Emotion Recognition from Text Using Semantic Labels and Separable Mixture Models," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 5, no. 2, pp. 165–181, 2006. Article (CrossRef Link).

[18] C. Hung and H. K. Lin, "Using objective words in sentiwordnet to improve word-of-mouth sentiment classification," *IEEE Intell. Syst.*, vol. 28, no. 2, pp. 47–54, 2013. Article (CrossRef Link).

[19] Y. Kim, S. R. Jeong, and I. Ghani, "Text Opinion Mining to Analyze News for Stock Market Prediction," *Int. J. Adv. Soft Comput. Its Appl.*, vol. 6, no. 1, 2014. Article (CrossRef Link)

[20] Z. Shi, H. Rui, and A. Whinston, "Content sharing in a social broadcasting environment: evidence from twitter," *Mis Q.*, vol. 38, no. 1, pp. 123–142, 2014. Article (CrossRef Link).

Yoosin Kim is a Big-data Analytics Visiting Professor in the University of Seoul. He received a Ph.D. with a research for Stock Index Prediction of News Big-data from Kookmin University in Seoul, Korea. He was a post-doctoral researcher in the University of Texas at Arlington, a data scientist at Accenture, and business analyst at SK. He has studied a fire-risk prediction model, Social Economic Index, a public service process mining methodology, and big data analytics.

Mingon Kang is an Assistant Professor in the Department of Computer Science at Kennesaw State University. His current research interests include Machine Learning, Bioinformatics, Big Data Analytics, and Computer Vision. Especially, he is focusing on developing novel computational methodologies for sparse learning, subspace learning, data integration, and personalized learning. He has published about 40 research papers in prestigious journals and conferences. Dr. Kang obtained his Ph.D and master degrees from University of Texas at Arlington in 2015 and 2010 respectively, and has a B.E. in Computer Engineering from Hanyang University in South Korea.

Seung Ryul Jeong is a Professor in the Graduate School of Business IT at Kookmin University, Korea. He holds a B.A. in Economics from Sogang University, Korea, an M.S. in MIS from University of Wisconsin, and a Ph.D. in MIS from the University of South Carolina, U.S.A. Professor Jeong has published extensively in the information systems field, with over 60 publications in refereed journals like Journal of MIS, Communications of the ACM, Information and Management, Journal of Systems and Software, among others.