ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2018 September: 24(3): 45~66 http://dx.doi.org/10.13088/jiis.2018.24.3.045

Online Document Mining Approach to Predicting Crowdfunding Success*

Suhyeon Nam
School of Management, Kyung Hee University
(gloomycloud@khu.ac.kr)

Yoonsun Jin
School of Management, Kyung Hee University
(dudnrha@khu.ac.kr)

Ohbyung Kwon
School of Management, Kyung Hee University
(obkwon@khu.ac.kr)

Crowdfunding has become more popular than angel funding for fundraising by venture companies. Identification of success factors may be useful for fundraisers and investors to make decisions related to crowdfunding projects and predict a priori whether they will be successful or not. Recent studies have suggested several numeric factors, such as project goals and the number of associated SNS, studying how these affect the success of crowdfunding campaigns. However, prediction of the success of crowdfunding campaigns via non-numeric and unstructured data is not yet possible, especially through analysis of structural characteristics of documents introducing projects in need of funding. Analysis of these documents is promising because they are open and inexpensive to obtain. We propose a novel method to predict the success of a crowdfunding project based on the introductory text. To test the performance of the proposed method, in our study, texts related to 1,980 actual crowdfunding projects were collected and empirically analyzed. From the text data set, the following details about the projects were collected: category, number of replies, funding goal, fundraising method, reward, number of SNS followers, number of images and videos, and miscellaneous numeric data. These factors were identified as significant input features to be used in classification algorithms. The results suggest that the proposed method outperforms other recently proposed, non-text-based methods in terms of accuracy, F-score, and elapsed time.

Key Words : Crowdfunding, Text analysis, Classification, Online data

Received : May 9, 2018 Revised : May 9, 2018 Accepted : July 14, 2018
Publication Type : Regular Paper Corresponding Author : Ohbyung Kwon

1. Introduction

Companies with insufficient capital to bring new products or services to market typically borrow money from financial institutions, stocks, or bonds (Myers & Majluf, 1984). However, it is difficult to obtain sufficient funds from financial institutions in the case of small-scale enterprises or venture startups due to their higher credit risk. As a starting point for such venture business entrepreneurs, business angels and venture capitalists may contribute funding. Recently, with

* This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2017S1A3A2066740)

the emergence of social media technology and SNS-based platforms, crowdfunding has received attention in academic research (Schwienbacher & Larralde 2010).

Crowdfunding, which evolved from the concept of crowdsourcing, refers to "an open call, essentially through the Internet, for the provision of financial resources either in the form of a donation or in exchange for some form of reward and/or voting rights in order to support initiatives for specific purposes" (Schwienbacher & Larralde 2010). It has also been defined as "the efforts by entrepreneurial individuals and groups – cultural, social, and for profit – to fund their ventures by drawing on relatively small contributions from a relatively large number of individuals using the internet, without standard financial intermediaries" (Mollick 2014).

Crowdfunding is used for a variety of projects and purposes related to entrepreneurship, business expansion, and ideas. It has attracted worldwide attention as a new funding method for new venture creators. To aid potential investors, it is important to identify the factors that will predict project success, which many studies have attempted to do. Online open documents that provide text introducing crowdfunding projects provide abundant information to aid in decision-making. In this research, we assess the possibility of project success through analysis of this introductory document (Du et al., 2015). Previous studies examined various factors such as project goals, duration, and categories that might influence the outcomes of fundraising campaigns (Cordova et al., 2015). Although one study utilized the textual analysis approach to predict the success of crowdfunding projects (Yuan et al., 2016), their method involved simple extraction of keywords from text.

The purpose of this study is to identify the factors influencing success of crowdfunding projects and to propose a novel method of predicting project success. We focus on reward-based crowdfunding (using platforms such as Eppela, Ulule, Starteed, Indiegogo, and Kickstarter), which is the most widely used type of crowdfunding. In addition, we analyze the unstructured documents introducing crowdfunding projects using a unique text analysis technique. Since the document introducing a crowdfunding project sends certain signals to potential investors, and successful crowdfunding is related to signals of the quality of the proposed project (Mollick, 2014), it is worthwhile to analyze this introductory document as a tool to predict whether and to what extent the project will be successful.

The remainder of this paper is organized as follows. In Section 2, we examine the success factors related to and predictions about crowdfunding and crowdfunding projects, citing previous studies on the text mining techniques to be utilized. In Section 3, we describe our research method. In Section 4, the analysis and results are described. Section 5 discusses the results of the analysis. Section 6 describes the conclusions of this study and raises possible issues for future research.

2. Crowdfunding

Crowdfunding derives from microfinance (Morduch, 1999) and crowdsourcing, in which the development process involves the sharing of ideas and feedback from the public (Poetz & Schreier, 2012; Elizabeth et al., 2012). Crowdfunding has become established as a unique method of funding on purpose-built platforms. Crowdfunding is defined as "open recruitment through the Internet as financial resources for donation or reward exchange, special purpose support" (Schwienbacher & Larralde, 2010), or “the act of collecting external funds from the massive public" (Belleflamme et al., 2010). Crowdfunding is also "a specific enterprise, cultural, social, and commercial group that excludes existing financial intermediaries and realizes creative projects or venture businesses through the relatively small amount of money raised by a large number of individuals from the Internet" (Howe, 2008).

As shown in Table 1, resources raised through crowdfunding may be divided into four types: donation, lending, reward-based, and equity, depending on the target of investment (Dushnitsky, 2013; Giudici et al., 2012; Leimeister, 2012). Donation crowdfunding, which occurs via the Internet, differs from traditional donation in that it goes through brokers. However, as with general donations, material rewards cannot be expected. Lending crowdfunding is funded by sponsors, each with a micro-loan to offer (Bruton et al., 2015). Sponsors receive interest insofar as they are contracted (Giudici et al., 2012). This is also called a P2P loan because it is an individual loan (Hermer, 2011). Reward-based crowdfunding is the most common form in recent times (non-monetary compensation) (Mollick, 2014). Reward-type crowdfunding allows sponsors to obtain products at a better price by pre-purchasing or reserving in advance (Hemer, 2011; Rothler & Wenzlaff, 2011). One of the largest and oldest crowdfunding platforms, Kickstarter, offers reward-type crowdfunding and is the most frequently analyzed type of crowdfunding (Frydrych et al., 2014; Kuppuswamy et al., 2015). Equity crowdfunding involves becoming a shareholder in return for support from project sponsors (Brem & Wassong, 2014). In most cases, the amount is less than would be received from a venture capitalist or angel investor (Belleflamme et al., 2014). In this paper, we focus on reward-based crowdfunding, which is the most common model of crowdfunding.

〈Table 1〉 Types of Crowdfunding

Type	Characteristics
Donation	Support for pure donation purposes without presupposing compensation
Lending	Support through micro-loans on the Internet involving interest payments
Reward-based	A sponsor funds the project of the venture company and is compensated in a form other than money
Equity	Support for the purpose of earning profit by acquiring equity proportionate to the amount of investment

As the importance of crowdfunding has increased, many studies have been conducted on the factors influencing the success of crowdfunding, including the target amount and duration of the fundraising. According to Mollick (2014), Zheng et al. (2014), Burtch et al. (2013), and Kuppuswamy et al. (2015), these two factors have a significant effect on the success of a crowdfunding project.

Other important factors are the creator's social network, the number of SNS friends, and the amount of SNS activity (Nevin et al., 2017). Crowdfunding becomes known mainly through SNS and is affected by the number of friends of the founder on SNS (Zheng et al., 2014; Mollick, 2014; Kromidha et al., 2016). Similarly, the extent to which founders are active in SNS activities influences project success (Song & Boeschoten, 2015).

Another factor is the length of the project introduction text and explanation of project risk. The number of words or lines determines the length of the text. In the studies of Lee and Shin (2014) and Koch et al. (2015), the length of the document proved to be a factor affecting the success of crowdfunding. A final influential factor is the number of updates to the project introduction, according to Kraus et al. (2016), Mollick (2014), Kuppuswamy et al. (2015), and Joenssen et al., (2014). The number of updates to the project introduction has been found to be an important factor in project success. Thus, a frequently updated introduction is vital to project success, informing the sponsors about the status of the project and communicating enthusiasm for the founder's project.

In addition, the amount of funds raised initially, the number of sponsors, and the cumulative recruitment amount are also important. According to the research of Zhang et al. (2012) and Colombo et al. (2015), the amount of funds collected and number of donors at the beginning of the project greatly affects the success of the project. According to Agrawal et al. (2011), the cumulative amount of funds affects subsequent fundraising and project success.

Lastly, the period between the end of the project and the described start of shipment has also been shown to be important. The shorter the period, the higher the likelihood of project success (Joenssen et al., 2014).

In this paper, the content of the project introduction is the main focus. According to one study, the probability of success is high when the words "most popular", "recently launched", and "ending soon" are included (Kuppuswamy et al., 2015). In addition, Mollick (2014) noted that as the number of grammatical errors increased, the probability of success decreased. Creative sustainability, project feasibility, and project creativity manifested in the project’s introduction text also influence the success of the project (Calic et al., 2016). Also important were having information about the situation, information about prototypes, perceived compassion and enthusiasm of the founders, the option of postponed payment in the text related to project risk, and mentioning the possibility of project failure (Koch et al.,

2016).

Although most studies focus on numerical factors such as the length of the introduction, we make use of project introduction texts, with their valuable information, using text mining techniques. Other studies rely on statistical expertise in analyzing documents related to the success of crowdfunding campaigns, which is very costly and knowledge intensive. Therefore, to mitigate these problems, some studies include a text analysis of the project introduction using text mining techniques. First, Du et al. (2015) examined all projects on Kickstarter from 2009 to November 2014. These projects were in progress, canceled, or pending projects, all with a target value of less than $100. Projects with an introduction length of less than 100 words were excluded. In total, 154,561 projects were investigated. Using the Gunning Fog Index to measure readability, and including the positive/negative word ratio of the whole project introduction as a new variable, that study predicted project success by conducting a regression analysis using the success factors of previous crowdfunding projects as control variables. However, that study used only two kinds of information in the introductory text: readability and emotional word count. The results showed a limit of 71%, which is relatively low.

The study of Yuan et al. (2016) focused on China's crowdfunding platforms, Zhongchou (www.zhongchou.com) and Dreamore (www.dreamore.com), targeting 500 of every 1,000 projects. They utilized topic modeling and the DC-LDA technique to extract topic features of the articles introducing each project to predict project success. They also used the Random Forest average with existing numerical data. However, in that study, 23,113 online news articles were collected and used to obtain topic features. In order to collect such a large amount of data, considerable time and effort is required compared to using crowdfunding project information. Therefore, in this study, we utilize the extracted topic feature.

Finally, the study of Mitra et al. (2014) included 45,815 projects on Kickstarter which has been issued untill February 6, 2012. Through the preprocessing and n-gram methods, they extracted specific sets of phrases associated with project success and failure. Then they predicted the success of each crowdfunding project using the logistic regression method with the set of extracted phrases as control variables and those associated with success factors of the existing crowdfunding project. However, other attributes of the introduction have not been examined.

3. Method

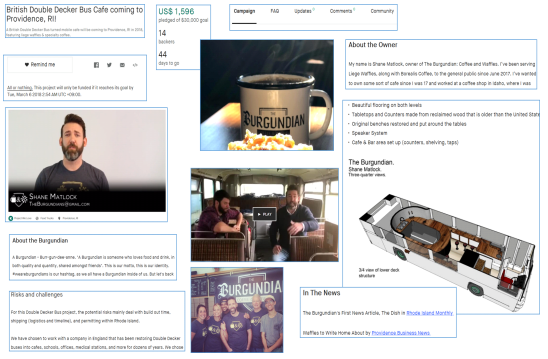
3.1 Inside the Crowdfunding Sites

Various sites around the world are devoted to crowdfunding. Kickstarter and Indiegogo are two representative examples, and Wadiz is the largest site in Korea. Upon entering this site, the introduction to each product is visible under the campaign or project title, forming a document with

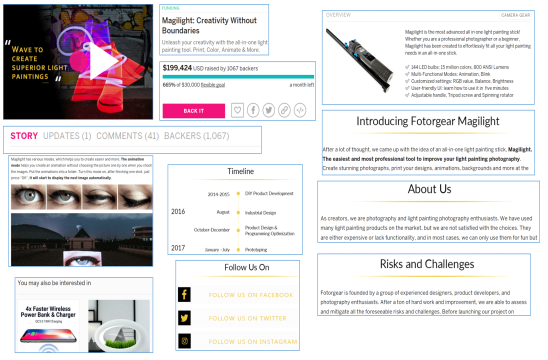
various types of information. As shown in Fig. 1, also visible are the title of the crowdfunding campaign, the target amount of the project and the amount achieved to date, the time remaining, videos, images, and text describing the product, an introduction to the venture company, and a number of comments, elements, and other types of related SNS and connection information. Table 2 summarizes the information contained in these three representative sites.

〈Table 2〉 Information Embedded in Three Crowdfunding Sites

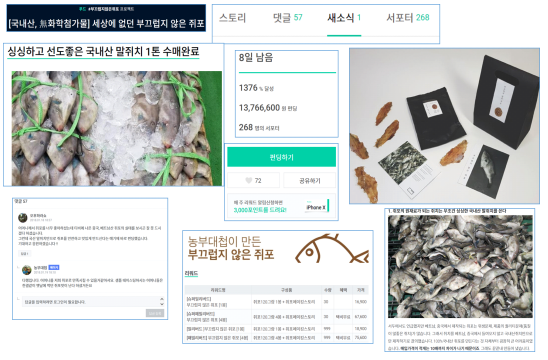
Name	Kickstarter	Indiegogo	Wadiz
Product category	O	O	O
Product theme			O
Theme	O	O	O
Number of replies/comments	O		O
Frequency of introduction updates			O
Fund recruitment start date	O	O	O
Fund recruitment end date	O	O	O
Target amount	O	O	O
Condition of reward	O	O	O
Number of SNS followers		O	O
Notified delivery due date		O	O
Image	O	O	O
Video	O	O	O
Number of past successful projects of founder	O	O	O
Number of past failure projects of founder	O	O	O
Number of projects invested by founders	O	O	O
Number of founders in a project	O	O	O
Fund recruitment achievement amount			O
Fund recruitment achievement rate			O
Abstract of the introduction part or overview	O	O	O
Introduction part	O	O	O
Warning part	O	O	O



(a) Kickstarter



(b) Indiegogo

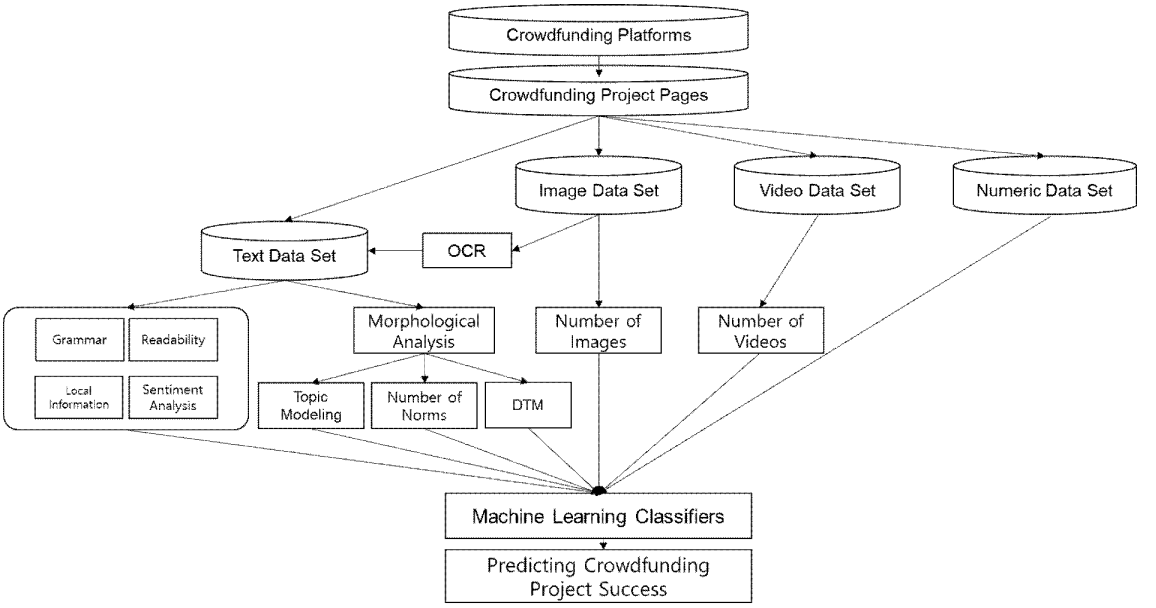


(c) Wadiz

〈Figure 1〉 Crowdfunding Sites

3.2 Overall System Framework

To predict the success of crowdfunding project using preestablished success factors and text mining techniques, we propose a framework like that illustrated in Fig. 2. First, we crawl the data required for analysis from the project page of the chosen crowdfunding platform. The collected data is then divided into text data, image data, video data, and numeric data. In addition, the image data is extracted through OCR and merged into the text data set. For the image and video data sets, the number of videos for each project is recorded and the data is stored as a numerical value. For the text data set, the number of misspellings, the readability index, inclusion of local information, and a score resulting from an emotional analysis are recorded for the entire document. Then, we perform a morphological analysis in which nouns are extracted to formalize the collected elements of the unstructured text. Then, we identify topic features for each project using topic modeling with the extracted nouns. In the next step, a Document Term Matrix (DTM) is generated with the nouns. The extracted values are used to learn and classify machine learning classifiers using Random Forest, SVM, and other classification algorithms. In addition, the performance of the learned model is evaluated using projects that were not terminated at the time of data collection.



〈Figure 2〉 System Framework

3.3 Preprocessing

The first task of preprocessing is to convert image information into text form through an OCR (optical character reader). An OCR is used because text data is often contained in an image file. Some projects even include all text in the image. Therefore, for more accurate analysis, it is necessary to extract the text contained in the image using an OCR.

Once all text information has been collected, a morphological analysis is performed. Since the main ideas of a document are included in the keywords, which are mainly nouns, the morphological analysis is crucial (Bouras & Tsogkas, 2008). Therefore, we extract nouns from the collected text data using Rhino 2.5.4 to perform a morphological analysis, saving them for later analysis.

Third, a spell check is performed on the collected text data. Spell checking is necessary to assess the degree to which spelling errors affect the success of the project (Mollick, 2014). The Hunspell R package is used along with the latest version of the related Korean dictionary, ko-aff-dic-0.7.0, as the Korean dictionary.

3.4 Readability Index Calculation

Since the readability of the project introduction affects the success of the crowdfunding project (Du et al., 2015), the readability index is also calculated for the collected text data in this study. Readability means the extent to which the content

of the text is easy to understand. In a previous study, the Gunning Fog formula was used to calculate readability. However, since this formula requires complex words and there are no proper complex words in Korean, we utilized another readability formula. Among a variety of readability formulas, according to Harrison's (1980) study, the Flesch (1948) readability formula, the readability graph of Fry (1968), and Dale-Chall (1948) are the most widely used formulas for readability calculation. However, since the Dale-Chall Formula requires complex words, we selected Flesch's readability formula, which is used by the Indiana State government in the USA and is also used in MS Word.

3.5 Sentiment Analysis

Emotional analysis was conducted on the introduction text for each crowdfunding campaign according to the protocol in another study (Du et al., 2015), which demonstrated that the degree of sensitivity of the text in the introduction affects the success of the project. Sentiment analysis, which is an important research area in the field of text mining, is an analytic technique that classifies how the speaker feels about the object (Liu, 2012; (Choi et al., 2017). A typical method of performing emotional analysis is to use an emotion dictionary (Cambria, 2013), which is a collection of emotional words. In this study, Sentiword Net, which is a commonly used emotion dictionary, is used.

3.6 Extracting Local Information

According to the findings of recent studies that founders and investors live in the same area and the physical distance between them affects the success of the funding campaign (Lin & Viswanathan, 2015; Burtch et al., 2013), we assumed that the inclusion of local information in the project introduction would affect project success. Therefore, we used the object name recognition technique to check whether text data (extracted nouns) contains local information. An entity name is a word or phrase that has a specific meaning in the document, such as a name, organization name, place name, time, date, and currency. Object name recognition methods include use of an entity name dictionary, the pattern-based method, and the machine learning method. Most studies use a machine learning technique. In this study, since only the place name and the name of the local government are used, the place name included in the article is recognized by the object name dictionary and the pattern-based method is used.

3.7 Topic Modeling

The topic of the post has been found to be an important factor influencing the success of crowdfunding (Yuan et al., 2016). Topic modeling is a widely used technique among text mining techniques for finding and classifying text topics. Latent Dirichlet Allocation (LDA) is one of the most popular topic modeling algorithms. Therefore, in this study, topic modeling using the

LDA algorithm with the text data extracted from the nouns was used to identify the topic features of each project introduction and considered as a factor for determining the success of crowdfunding campaigns.

3.8 DTM (TF-IDF)

To analyze unstructured text, it is necessary to formalize the data. To do this, a document-term matrix (DTM) was created using noun-extracted text data. When creating the DTM, we removed the stopwords from the unstructured text, preprocessed them using feature selection, and used the TF-IDF (Term Frequency-Inverse Document Frequency) weight, which is commonly used as the weight of nouns in the DTM. In addition, since the high-dimensionality problem leads to longer learning times and less accuracy if all nouns remain in the DTM, feature selection is performed based on the Chi-squared test.

3.9 Miscellaneous Data

As with other data, the number of images and videos inserted in the document is determined. In addition, the fundraising period, the period between the fundraising start and end dates, is measured in days. Then, the period between the fundraising end date and the noticed delivery date is set as the period up to the scheduled delivery date. For the start date, we included the year as the project year, but did not include the end date as a variable. We counted the reward types to determine reward diversity. We included the introductory article and the risk-related text in the noun capital variable. Finally, for the fundraising target and the number of SNS followers, in cases with large deviations, the scale was reduced by the log function. Other collected data was used as-is.

Table 3 summarizes the information obtained through the above text analysis procedure.

〈Table 3〉 Information Obtained through Text Analysis

Feature	Type	Feature	Type
Product category	Nominal	Inclusion of local information	True/False
Number of replies	Numeric	Number of typos	Numeric
Frequency of introduction updates	Numeric	Number of past projects of founder	Numeric
Fund recruitment target amount	Numeric	Number of past successful projects of founder	Numeric
Fund recruitment year	Numeric	Number of past failure projects of founder	Numeric
Fund recruitment period	Numeric	Number of projects invested in by founders	Numeric
Fund recruitment type	Nominal	Number of nouns in the introduction about the project	Numeric
Fund recruitment end date - Duration of delivery start date	Numeric	Number of nouns in the warning statement about the project	Numeric
Variety of rewards	Numeric	Topic feature	Nominal
Number of SNS followers	Numeric	Terms (TF-IDF)	Numeric
Level of readability	Numeric		

3.10 Predicting Project Funding Results

To derive an optimal prediction function, all data used in the experiment were randomly divided into a training set and test set at a ratio of 7:3 using the holdout method. Then, learning was performed by repeating the experiment 100 times to obtain the confusion matrix, and the performance of the algorithm was evaluated. The F1 score, which is a harmonic average of precision and recall, and elapsed time were used as the metrics for the evaluation. We then selected the best classification algorithm through these methods and metrics, analyzing the performance thereof, classifying the projects that were not terminated at recruitment time, and finally confirming the performance thereof.

Finally, classification learning proceeds with the input features used in this study. The final features used in learning are shown in Table 3. The classification uses C5.0, CART, DNN, k-NN, logistic regression, naïve Bayes, Random Forest, and SVM. Parameter values are optimized using a tuning function or heuristic method. In other words, we evaluate the performance through repeated experiments to find the most suitable algorithm.

4. Experiment

4.1 Data Collection and Simple Statistics

In this study, we selected Wadiz as one of the domestic crowdfunding platforms for data collection. Wadiz is the largest crowdfunding platform in Korea, with a 60% share of the crowdfunding market, a membership of 600,000, and an annual investment of KRW 15 billion (Financial Services Commission, 2017). It is an example of a rewards-type crowdfunding platform. Therefore, in this study, we analyzed rewards-type crowdfunding. Information with corresponding data types collected from crowdfunding projects is shown in Table 4.

From January 27, 2014 to October 31, 2017, we collected data for all rewards-type funding projects registered at Wadiz. Data for a total of 2203 projects were collected, of which 112 were not completed at the time of collection and 2091 were completed. Of the 2091 completed projects, 1196 were successful, accounting for 57.2% of the total, and 895 projects failed, accounting for 42.8% of the total. The average amount raised was about 7,300 USD, and the average goal achievement rate was 220.09%. By category, there were 355 projects in the technology category (17% of the total), 209 in the fashion/beauty category, 194 in the food category (10%), 473 in the design product category (23%), 4 in other categories (0.4%), 110 in publishing (5%), and so forth. By year, the average number of comments gradually increases to 185 in 2014, 493 in 2015, 697 in 2016, and 716 in 2017. The average total number of comments is 78.4, and the average update rate is 4.5.

〈Table 4〉 Data Collected from a Crowdfunding Project

Name	Type
Product category	Nominal
Reward	Nominal
Number of replies	Numeric
Frequency of introduction updates	Numeric
Fund recruitment start date	Date
Fund recruitment end date	Date
Target amount	Numeric
Condition of reward	1 (unconditional), 2 (if successful)
Number of SNS followers	Numeric
Notified delivery due date	Date
Image	Binary
Video	Binary
Number of past successful projects of founder	Numeric
Number of past failure projects of founder	Numeric
Number of projects invested in by founders	Numeric
Number of founders in a project	Numeric
Fund recruitment achievement amount	Numeric
Fund recruitment achievement rate	Numeric
Theme	Text
Abstract of the introduction	Text
Introduction	Text
Warning	Text

4.2 Regression Analysis

A regression analysis was performed to predict if a project's crowdfunding campaign would result in success or failure. In the regression analysis, 111 projects with fewer than 100 nouns extracted from 2091 projects were judged to have insufficient meaningful text data; the remaining 1980 projects were included in the regression analysis. Dependent variables were coded as 1 and 0, which indicate project success and failure, respectively. The independent variables were the input features, which were calculated by analyzing the collected text while excluding the number of comments and the number of project introduction updates.

The results of the regression analysis (Table 5) indicate significant effects of *Log_goal*, *Type*, *Log_sns*, *Neg*, *Past_project*, *Past_success_project*, *Fund_project*, and *N_num* on product success, where *Log_goal* indicates the project fundraising goal, type is the project's reward type, *Variety_of_gift* is the diversity of the rewards, and *Log_sns* is the variable that logs the number of a founder's SNS followers. *Neg* is the number of words with negative emotions in the text, and *Past_project* is the number of projects the founder has done in the past. *Past_success_project* is the number of projects with which the founder has been successful in the past. *Fund_project* refers to whether the founder has funded another project. *N_num* is the number of extracted nouns. Topic is the topic of the introductory article obtained through topic modeling. The adjusted R-squared value of the overall regression equation for the full model was 0.2349.

In addition, we considered two competitive models, reduced Models 1 and 2, to search for a better model than the full model. Reduced Model

〈Table 5〉 Regression Analysis for Project Success

	Full model	Reduced model 1	Reduced model 2
(Intercept)	0.589	0.049 *	0.466
Log_goal	0.001 ***	0.001 ***	0.001 ***
Start_date	0.567	0.044 *	0.450
Terms	0.181	0.226	0.114
Type	0.001 ***	0.001 ***	0.001 ***
Delivery_term	0.841	0.755	0.901
Variety_of_gift	0.007 **	0.001 ***	0.006 **
Log_sns	0.001 ***	0.001 ***	0.001 ***
Image_num	0.121	0.001 ***	0.074
Video_num	0.731	0.750	0.697
Past_project	0.001 ***	0.001 ***	0.001 ***
Past_success_project	0.001 ***	0.001 ***	0.001 ***
Fund_project	0.033 *	0.025 *	0.029 *
Topic	0.001 ***		0.001 ***
Pos	0.590		0.590
Neg	0.005 **		0.005 **
Senti	0.347		0.382
Location	0.921		0.887
Readability_num	0.358		0.394
Wrong_word_num	0.106		0.088
N_num	0.001 ***		0.001 ***
Risk_word_num	0.774		0.799
Tech	0.740	0.939	
Fashion/Beauty	0.374	0.490	
Food	0.705	0.996	
Design	0.475	0.493	
Web tune	0.011 *	0.022 *	
Game	0.503	0.749	
Publication	0.082	0.196	
Popular art	0.725	0.801	
Public project	0.564	0.579	
Travel/Leisure	0.468	0.525	
P-value	2.2e-16	2.2e-16	2.2e-16
R^2	0.2469	0.2258	0.2403
Adjusted R^2	0.2349	0.2161	0.2321
N	1980	1980	1980

* $p<0.05$; ** $p<0.01$; *** $p<0.001$

〈Table 6〉 Regression Analysis for Amount Achieved

Dependent variable	Log_Achievement amount		
	Full model	Reduced model 1	Reduced model 2
(Intercept)	0.064	0.01 **	0.028 *
Log_goal	0.001 ***	0.001 ***	0.001 ***
Start_date	0.062	0.009 **	0.028 *
Terms	0.137	0.138	0.135
Type	0.001 ***	0.001 ***	0.001 ***
Delivery_term	0.709	0.779	0.697
Variety_of_gift	0.001 ***	0.001 ***	0.001 ***
Log_sns	0.001 ***	0.001 ***	0.001 ***
Image_num	0.002 **	0.001 ***	0.001 ***
Video_num	0.976	0.985	0.945
Past_project	0.001 ***	0.001 ***	0.001 ***
Past_success_project	0.001 ***	0.001 ***	0.001 ***
Fund_project	0.518	0.54	0.665
Topic	0.025 *		0.027 *
Pos	0.164		0.163
Neg	0.149		0.175
Senti	0.710		0.785
Location	0.142		0.222
Readability_num	0.803		0.697
Wrong_word_num	0.138		0.098
N_num	0.001 ***		0.001 ***
Risk_word_num	0.630		0.516
Tech	0.583	0.755	
Fashion/Beauty	0.319	0.41	
Food	0.651	0.806	
Design	0.347	0.379	
Web tune	0.001 ***	0.001 ***	
Game	0.338	0.395	
Publication	0.271	0.367	
Popular art	0.186	0.181	
Public project	0.511	0.524	
Travel/Leisure	0.973	0.968	
P-value	2.2e-16	2.2e-16	2.2e-16
R^2	0.1979	0.1851	0.1864
Adjusted R^2	0.1851	0.176	0.1776
N	1980	1980	1980

* p<0.05; ** p<0.01; *** p<0.001

1 excluded variables generated from the text analysis, while reduced Model 2 does not consider the type of products and services introduced in the corresponding crowdfunding project. The results suggest that information extracted from introductions about crowdfunding projects by text analysis significantly helps in predicting the success of the project. Furthermore, for reduced Model 2, the differences in the adjusted R-squared value be-tween the full model and reduced Model 2 are very similar (0.2349 and 0.2321).

In the second regression model, the project funding achievement was set as the dependent variable. The independent variables were the same as in the first regression model. The results (Table 6) suggest that overall project success is consistent with the regression results, but the Fund_project variable is not significant; however, the number of images is significant.

4.3 Predicting Crowdfunding Success

The performance of learned classifiers is evaluated by measuring accuracy, the F-score, and elapsed time. The results are shown in Table 7. For the classifier using text information, the Random Forest algorithm showed the best results (overall accuracy 0.878446, F-score 0.816204, elapsed time 6.7 seconds). There was no significant difference among classifiers in determining the learned model. The Random Forest method was the highest of the classifiers that did not use text information. However, accuracy and F-score showed a difference of 0.16 or more between when text information was used and not used. Therefore, we conclude that higher classification accuracy can be obtained when using text information. Consequently, the proposed method outperforms that any previous study, with a score of 78.93% (Koch et al., 2016).

Table 8 shows the results of the classification performance test for the 112 projects finished after the data collection time with the learned model. As in the previous test, the Random Forest method showed the highest performance in terms of accuracy (0.8235) and F-score (0.8594).

〈Table 7〉 Comparison of Performance of Crowdfunding Success Judgment by Text Analysis

Algorithms	Including Text Analysis			Excluding Text Analysis		
	Accuracy	F-score	Elapsed time (sec)	Accuracy	F-score	Elapsed time (sec)
C5.0	0.8428	0.7583	3.2	0.6877	0.6286	0.2
K-NN	0.7166	0.5837	0.5	0.6591	0.5940	0.2
Logistic Regression	0.5711	0.3214	0.6	0.5529	0.2422	0.2
Naïve Bayes	0.5813	0.3396	0.7	0.5727	0.0719	0.2
Random Forest	0.8784	0.8162	6.7	0.7156	0.6487	0.5
SVM	0.8653	0.8124	48.3	0.6928	0.6745	1.0

〈Table 8〉 Comparison of Classification Algorithms

Algorithms	Accuracy	F-score	Elapsed time (sec)
C5.0	0.5540	0.6786	0.1
K-NN	0.6735	0.7826	0.1
Logistic Regression	0.5423	0.6594	0.1
Naïve Bayes	0.6723	0.6422	0.1
Random Forest	0.8235	0.8594	0.1
SVM	0.8023	0.7923	0.2

5. Discussion

5.1 Implications

The academic implications of this study are as follows. First, we analyzed various attributes of the introduction to crowdfunding projects. Previous studies of the role of the introduction in predicting success used only one or two attributes from the introduction page. We examined topic, readability, and sentiment simultaneously, applying unstructured textual analysis methods to improve prediction accuracy and comparing the influences of various properties.

This study was the first to use DTM derived from the introduction to predict the success of crowdfunding projects. In this paper, we propose a method to estimate the distribution of all terms included in the introduction about a given crowdfunding project, making extensive use of the entire contents of the text and improving prediction accuracy.

The method proposed herein is better than those utilized in recent studies, which predict the success of crowdfunding campaigns based on unstructured online data. The most prominent example is Yuan et al. (2016), in which factors are suggested that influence success using topic modeling and keyword analysis. However, our study includes more relevant input features and contextual information, including factors obtained through topic modeling; are method is superior because making a prediction occurs faster.

This study has the following practical implications. First, the results revealed the factors affecting the success of crowdfunding projects. In addition, we suggested ways for entrepreneurs to increase the possibility of project success before starting their fundraising campaigns(Lee et al., 2016). Furthermore, unlike in previous research, we present meaningful guidelines regarding the project introduction. In fact, most of the information that crowdfunding investors see is provided on the project introduction page. Information asymmetry is a problem for investors initially because they cannot know about the end product or the result of the crowdfunding campaign. Therefore, from the viewpoint of signaling theory, signals must be sent to investors through the introduction. In this study, we provide guidelines for optimal signaling to investors.

5.2 Conclusion

In this paper, we proposed a method of predicting the success of a crowdfunding project

using classification algorithms to determine the factors influencing success of reward-type crowdfunding projects on a large crowdfunding platform. We focused on reward-type crowdfunding projects listed on Wadiz and utilized various text analysis techniques such as emotional analysis, object name recognition, and readability index to evaluate unstructured data. The results of the analysis and the numeric information obtained from the crowdfunding project were used as regressors to determine the degree of influence of these various factors. Using various classification algorithms for several variables, we treated the nouns in the introductions as features. We evaluated the performance of the finished projects after the data analysis.

This study has some limitations that could be improved in future research. First, we utilized a single Korean crowdfunding platform called Wadiz. Although Wadiz is the largest crowdfunding platform in Korea, it still has the limitation of being a single platform. Therefore, in future studies, data should be collected from various platforms for experimentation, because the characteristics of the platform may influence the analysis. Secondly, we utilized OCR to extract text data from images. However, the text could not be fully extracted due to the limitations in performance of the OCR. Therefore, it is possible that the text information is not completely secure, which may cause distortion of the analysis results. In future studies, improved OCR performance should be a focus. Finally, this study only examined the project introduction, but the contents of the project may also affect its success. Therefore, analysis in future studies through technology such as image processing may provide more accurate results.

References

Agrawal, A. K., C. Catalini, and A. Goldfarb, *The geography of crowdfunding*, National Bureau of Economic Research Working Paper No.16820, 2011. Available at: http://www.nber.org/papers/w16820 (Accessed 8 May, 2018).

Beier, M., and K. Wagner, "Crowdfunding Success: A Perspective from Social Media and E-Commerce," *Proceedings of the 36th International Conference on Information Systems*, (2015).

Belleflamme, P., T. Lambert, and A. Schwienbacher, "Crowdfunding: An industrial organization perspective," *Proceedings of Workshop on Digital Business Models: Understanding Strategies*, (2010), 25-26.

Belleflamme, P., T. Lambert, and A. Schwienbacher, "Crowdfunding: Tapping the right crowd," *Journal of business venturing*, Vol.29, No.5(2014), 585-609.

Bouras, C., and V. Tsogkas, "Improving Text Summarization Using Noun Retrieval Techniques," Proceedings of the 12th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, (2008), 593-600.

Brem, A. and N. Wassong, "Wer investiert warum? Eine analyse von investmententscheidungen

bei crowdfunding-projekten," *ZfKE–Zeitschrift für KMU und Entrepreneurship*, Vol.62, No.1(2014), 31-55.

Bruton, G., S. Khavul, D. Siegel, and M. Wright, "New financial alternatives in seeding entrepreneurship: Microfinance, crowdfunding, and peer-to-peer innovations," *Entrepreneurship Theory and Practice*, Vol.39, No.1(2015), 9-26.

Burtch, G., A. Ghose, and S. Wattal, "An empirical examination of the antecedents and consequences of contribution patterns in crowd-funded markets," *Information Systems Research*, Vol.24, No.3(2013), 499-519.

Burtch, G., A. Ghose, and S. Wattal, "Cultural differences and geography as determinants of online pro-social lending," *MIS Quarterly,* Vol.38, No.3(2013), 773-794.

Calic, G. and E. Mosakowski, "Kicking off social entrepreneurship: how a sustainability orientation influences crowdfunding success," *Journal of Management Studies*, Vol.53, No.5(2016), 738-767.

Cambria, E., B. Schuller, Y. Xia, and C. Havasi, "New avenues in opinion mining and sentiment analysis," *IEEE Intelligent Systems*, Vol.28, No.2(2013), 15-21.

Choi, S. Lee, J., Kwon, O., "Financial Fraud Detection using Text Mining Analysis against Municipal Cybercriminality," *Journal of Information Systems,* Vol.23, No.3(2017), 119-138.)

(최석재, 이중원, 권오병 "지자체 사이버 공간 안전을 위한 금융사기 탐지 텍스트 마이닝 방법," *지능정보연구*, Vol.23, No.3(2017), 119-138.)

Colombo, M. G., C. Franzoni, and C. Rossi-Lamastra, "Internal social capital and the attraction of early contributions in crowdfunding," *Entrepreneurship Theory and Practice*, Vol.39, No.1(2015), 75-100.

Cordova, A., J. Dolci, and G. Gianfrate, "The determinants of crowdfunding success: evidence from technology projects," *Procedia-Social and Behavioral Sciences*, Vol.181, (2015), 115-124.

Dale, E., and J. S. Chall, "A formula for predicting readability: Instructions," *Educational research bulletin*, (1948), 37-54.

Du, Q., W. Fan, Z. Qiao, G. Wang, X. Zhang, and M. Zhou, "Money Talks: A Predictive Model on Crowdfunding Success Using Project Description," *Proceedings of the 21st Americas Conference on Information Systems*, (2015), 1-8.

Dushnitsky, G. and D. Marom, "Crowd monogamy," *London Business School Review*, Vol.24, No.4(2013), 24-26.

Flesch, R., "A new readability yardstick," *Journal of applied psychology*, Vol.32, No.3(1948), 221.

Fry, E., "A readability formula that saves time," *Journal of reading*, Vol.11, No.7(1968), 513-578.

Frydrych, D., A. J. Bock, T. Kinder, and B. Koeck, "Exploring entrepreneurial legitimacy in reward-based crowdfunding," *Venture Capital*, Vol.16, No.3(2014), 247-269.

Gerber, E. M., J. S. Hui, and P. Y. Kuo, "Crowdfunding: Why people are motivated to post and fund projects on crowdfunding platforms," *Proceedings of the International Workshop on Design, Influence, and Social Technologies: Techniques, Impacts and*

Ethics, (2012).

Giudici, G., R. Nava, C. Rossi Lamastra, and C. Verecondo, "Crowdfunding: The new frontier for financing entrepreneurship?," *SSRN Working Paper*, (2012).

Harrison, C. J. O., *Readability in the classroom*. Cambridge University Pres, Cambridge, 1980.

Hemer, J., *A snapshot on crowdfunding,* Working papers firms and region, 2011. Available at: http://hdl.handle.net/10419/52302 (Accessed 7 May, 2018).

Howe, J., *Crowdsourcing—How the Power of the Crowd Is Driving the Future of Business*, Random House Business, New York, 2008.

Joenssen, D., A. Michaelis, and T. Müllerleile, "A Link to New Product Preannouncement: Success Factors in Crowdfunding," *SRN Working Paper*, (2014).

Koch, J. A. and M. Siering, "Crowdfunding success factors: the characteristics of successfully funded projects on crowdfunding platforms," *Proceedings of the 23rd European Conference on Information Systems*, (2015).

Koch, J. A. and Q. Cheng, "The Role of Qualitative Success Factors in the Analysis of Crowdfunding Success: Evidence from Kickstarter," *Proceedings of the 20th Pacific Asia Conference on Information Systems*, (2016).

Kraus, S., C. Richter, A. Brem, C. F. Cheng, and M. L. Chang, "Strategies for reward-based crowdfunding campaigns," *Journal of Innovation & Knowledge*, Vol.1, No.1(2016), 13-23.

Kromidha, E. and P. Robson, "Social identity and signaling success factors in online crowdfunding," *Entrepreneurship & Regional Development*, Vol.28, No.9-10(2016), 605-629.

Kuppuswamy, V. and B. L. Bayus, "Crowdfunding creative ideas: The dynamics of project backers in Kickstarter," *UNC Kenan-Flagler Research Paper*, No.2013-15(2013), 605～629. Available at https://ssrn.com/abstract=2234765 (Accessed 7 May, 2018).

Lee, H. Jin, Y., Kwon, O., "Investigating the Impact of Corporate Social Responsibility on Firm's Short- and Long-Term Performance with Online Text Analytics," *Journal of Information Systems,* Vol.22, No.2(2016), 13-31.)

(이희승, 진윤선, 권오병 "온라인 텍스트 분석을 통해 추정한 기업의 사회적책임 성과가 기업의 단기적 장기적 성과에 미치는 영향 분석," *지능정보연구*, Vol.22, No.2(2016), 13-31.)

Lee, J. and H. D. Shin, "The Relationship between the Information Posted on the Web and the Success of Funding in Crowdfunding Site," *The Journal of the Korea Contents Association*, Vol.14, No.6(2014), 54-62.

(이정은, 신형덕, "크라우드펀딩 사이트의 게시글 정보가 펀딩 성공에 미치는 영향," *한국콘텐츠학회논문지*, Vol.14, No.6(2014), 54-62.)

Leimeister, J. M., "Crowdsourcing," *Controlling & Management*, Vol.56, No.6(2012), 388-392.

Lin, M. and S. Viswanathan, "Home bias in online investments: An empirical study of an online crowdfunding market," *Management Science*, Vol.62, No.5(2015), 1393-1414.

Liu, B., "Sentiment analysis and opinion mining,"

Synthesis lectures on human language technologies, Vol.5, No.1(2012), 1-167.

Mitra, T, and E. Gilbert, "The language that gets people to give: Phrases that predict success on kickstarter," *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, (2014), 49-61.

Mollick, E., "The dynamics of crowdfunding: An exploratory study," *Journal of business venturing*, Vol.29, No.1(2014), 1-16.

Morduch, J.," The microfinance promise," *Journal of economic literature*, Vol.37, No.4(1999), 1569-1614.

Myers, S. C. and N. S. Majluf, "Corporate financing and investment decisions when firms have information that investors do not have," *Journal of financial economics*, Vol.13, No.2(1984), 187-221.

Nevin, S., R. Gleasure, P. O'Reilly, J. Feller, S. Li, and J. Cristoforo, "Social Identity and Social Media Activities in Equity Crowdfunding," *Proceedings of the 13th International Symposium on Open Collaboration*, (2017).

Poetz, M. K. and M. Schreier, "The value of crowdsourcing: can users really compete with professionals in generating new product ideas?," *Journal of product innovation management*, Vol.29, No.2(2012), 245-256.

Röthler, D. and K. Wenzlaff, "Crowdfunding schemes in Europe," *EENC Report*, No.9 (2011).

Schwienbacher, A. and B. Larralde, *Crowdfunding of small entrepreneurial ventures*, Cumming, D.J. (Ed.), The Oxford Handbook of Entrepreneurial Finance, Oxford University Press, Oxford, 2010.

Song, Y., and R. van Boeschoten, Success factors for Crowdfunding founders and funders. *arXiv preprint arXiv:1503.00288*, 2015.

Yuan, H., R. Y. Lau, and W. Xu, "The determinants of crowdfunding success: A semantic text analytics approach," *Decision Support Systems*, Vol.91, (2016), 67-76.

Zhang, J. and P. Liu, "Rational herding in microloan markets," *Management science*, Vol.58, No.5(2012), 892-912.

Zheng, H., D. Li, J. Wu, and Y. Xu, "The role of multidimensional social capital in crowdfunding: A comparative study in China and US," *Information & Management*, Vol.51, No.4(2014), 488-496.

국문요약

온라인 문서 마이닝 접근법을 활용한 크라우드펀딩의 성공여부 예측 방법

남수현* · 진윤선* · 권오병**

크라우드펀딩(Crowdfunding)은 최근 벤처 기업의 기금 모금을 위한 엔젤 기금보다 인기가 있다. 이에 따라 크라우드펀딩의 성공 요인을 파악하는 것은 기금 조성자 및 투자자로 하여금 크라우드펀딩 프로젝트와 관련된 효과적 의사결정을 내리기 위해 크라우드펀딩 성공 여부를 선험적으로 예측하는데 유용할 것이다. 이에 최근까지 프로젝트의 목표 및 관련 SNS의 수와 같은 몇 가지 수치적 요인을 독립변인으로 제안하여 이들이 크라우드펀딩 캠페인의 성공에 어떤 영향을 미치는지 등이 연구되어오고 있었다. 그러나 수치가 아닌 비정형 데이터를 통한 크라우드펀딩 캠페인의 성공에 대한 예측은 거의 이루어진 바 없으며, 특히 프로젝트를 소개하는 문서에 대한 특성 분석을 통해 해당 프로젝트의 성공 여부를 예측하려는 연구는 아직 이루어지지 않았다. 사실 프로젝트를 소개하는 문서는 공개되어 있어 확보에 드는 비용이 적게 들기 때문에 매우 유용하다. 따라서 본 연구의 목적은 Wadiz 등 온라인상으로 공개되어 있는 프로젝트에 대한 소개 문서를 기반으로 크라우드펀딩 프로젝트의 성공을 예측하는 새로운 방법을 제안하는 것이다. 제안된 방법의 성능을 테스트하기 위해, 본 연구에서는 1,980개의 실제 크라우드펀딩 프로젝트와 관련된 텍스트를 수집하고 경험적으로 분석했다. 텍스트 데이터 세트에서 카테고리, 응답 수, 자금 조달 목표, 기금 모금 방법, 보상, SNS 추종자 수, 이미지 및 비디오 수 및 기타 숫자 데이터와 같은 프로젝트에 대한 세부 정보를 수집하였다. 분석 결과 이러한 요인들은 분류 알고리즘에서 분류 성능을 제고하는데 의미 있는 변인으로 확인되었다. 즉, 제안된 방법이 최근에 제안된 비정형 텍스트 기반 방법보다 정확도나 F-점수 및 수행 경과 시간에서 성능이 우수하였다.

주제어 : 크라우드펀딩, 텍스트분석, 분류 알고리즘, 온라인 문서

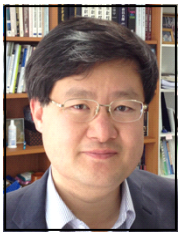
논문접수일 : 2018년 5월 9일 논문수정일 : 2018년 5월 9일 게재확정일 : 2018년 7월 14일
원고유형 : 일반논문 교신저자 : 권오병

* School of Management, Kyung Hee University
** Corresponding Author: Ohbyung Kwon
School of Management, Kyung Hee University
26 Kyung Hee Dae Ro, Dongdaemun-gu, Seoul 130-722, Korea
Tel: +82-961-2148, Fax: +82-961-0515, E-mail: obkwon@khu.ac.kr

저 자 소 개

남 수 현
경희대학교 일반대학원 경영학과에서 빅데이터 경영 전공으로 석사학위를 취득하였다. 주요관심분야는 Data Mining, Big data analysis, Text Mining, Algorithm Implementation 등이다.

진 윤 선
숙명여자대학교에서 e비즈니스학으로 석사학위를 취득하고, 현재 경희대학교에서 빅데이터경영 전공으로 박사과정에 재학 중이다. 주요 관심분야는 빅데이터, 공공데이터, 데이터마아닝, 텍스트마이닝 등이다.



권 오 병
현재 경희대학교 경영학과 교수로 재직 중이다. 1988년 서울대학교 경영학과(경영학사), 1990년 한국과학기술원 경영과학과(공학석사), 1995년 한국과학기술원 경영과학과(공학박사)를 졸업하였다. 관심분야는 빅데이터분석, 사물인터넷, 의사결정지원시스템 등이다.