ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2016 June: 22(2): 33~56 http://dx.doi.org/10.13088/jiis.2016.22.2.033

Bankruptcy Prediction Modeling Using Qualitative Information Based on Big Data Analytics*

Nam-ok Jo
School of Business, Ewha Womans University
(namok.jo@gmail.com)

Kyung-shik Shin
School of Business, Ewha Womans University
(ksshin@ewha.ac.kr)

Many researchers have focused on developing bankruptcy prediction models using modeling techniques, such as statistical methods including multiple discriminant analysis (MDA) and logit analysis or artificial intelligence techniques containing artificial neural networks (ANN), decision trees, and support vector machines (SVM), to secure enhanced performance. Most of the bankruptcy prediction models in academic studies have used financial ratios as main input variables. The bankruptcy of firms is associated with firm's financial states and the external economic situation. However, the inclusion of qualitative information, such as the economic atmosphere, has not been actively discussed despite the fact that exploiting only financial ratios has some drawbacks. Accounting information, such as financial ratios, is based on past data, and it is usually determined one year before bankruptcy. Thus, a time lag exists between the point of closing financial statements and the point of credit evaluation. In addition, financial ratios do not contain environmental factors, such as external economic situations. Therefore, using only financial ratios may be insufficient in constructing a bankruptcy prediction model, because they essentially reflect past corporate internal accounting information while neglecting recent information.

Thus, qualitative information must be added to the conventional bankruptcy prediction model to supplement accounting information. Due to the lack of an analytic mechanism for obtaining and processing qualitative information from various information sources, previous studies have only used qualitative information. However, recently, big data analytics, such as text mining techniques, have been drawing much attention in academia and industry, with an increasing amount of unstructured text data available on the web. A few previous studies have sought to adopt big data analytics in business prediction modeling. Nevertheless, the use of qualitative information on the web for business prediction modeling is still deemed to be in the primary stage, restricted to limited applications, such as stock prediction and movie revenue prediction applications. Thus, it is necessary to apply big data analytics techniques, such as text mining, to various business prediction problems, including credit risk evaluation. Analytic methods are required for processing qualitative information represented in unstructured text form due to the complexity of managing and processing unstructured text data.

This study proposes a bankruptcy prediction model for Korean small- and medium-sized construction firms using both quantitative information, such as financial ratios, and qualitative information acquired from economic news articles. The performance of the proposed method depends on how well information types are transformed from qualitative into quantitative information that is suitable for incorporating into the bankruptcy prediction model. We employ big data analytics techniques, especially text mining, as a mechanism for processing qualitative information. The sentiment index is provided at the industry level by extracting from a large amount of text data

* This work was supported by the National Research Foundation of Korea Grant funded by the Korean government (NRF-2013S1A3A2054667).

to quantify the external economic atmosphere represented in the media. The proposed method involves keyword-based sentiment analysis using a domain-specific sentiment lexicon to extract sentiment from economic news articles. The generated sentiment lexicon is designed to represent sentiment for the construction business by considering the relationship between the occurring term and the actual situation with respect to the economic condition of the industry rather than the inherent semantics of the term.

The experimental results proved that incorporating qualitative information based on big data analytics into the traditional bankruptcy prediction model based on accounting information is effective for enhancing the predictive performance. The sentiment variable extracted from economic news articles had an impact on corporate bankruptcy. In particular, a negative sentiment variable improved the accuracy of corporate bankruptcy prediction because the corporate bankruptcy of construction firms is sensitive to poor economic conditions. The bankruptcy prediction model using qualitative information based on big data analytics contributes to the field, in that it reflects not only relatively recent information but also environmental factors, such as external economic conditions.

Key Words : Bankruptcy Prediction, Big Data Analytics, Text Mining, Sentiment Analysis, Artificial Neural Networks

Received : May 3, 2016 Revised : May 31, 2016 Accepted : June 13, 2016
Publication Type : Regular Paper Corresponding Author : Kyung-shik Shin

1. Introduction

Bankruptcy prediction has steadily been a researched issue in the accounting and finance fields since the late 1960s. Many researchers have developed a more robust bankruptcy prediction model in terms of classification accuracy. While early studies adopted statistical techniques such as multiple discriminant analysis (MDA) (Altman, 1968) and logit analysis (Hamer, 1984; Ohlson, 1980), later studies adopted artificial intelligence (AI) approaches such as artificial neural networks (ANN) (Fletcher and Goss, 1993; Leshno and Spector, 1996; Odom and Sharda, 1990; Tam and Kiang, 1992), decision trees (Shaw and Gentry, 1990), and support vector machines (SVM) (Shin et al., 2005) as substitute methodologies for business prediction problems.

Bankruptcy of firms is associated with a firm's financial state and the external economic situation. However, studies on the utilization of qualitative information for a bankruptcy prediction model have yet to be conducted, despite the steady flow of study on construction of bankruptcy prediction models in terms of modeling techniques, such as statistical methods and artificial intelligence techniques. Even though using only financial ratios is known to be insufficient for bankruptcy prediction modeling, research on development of bankruptcy prediction models has mainly used only financial ratios as input variables.

Bankruptcy prediction models relying only on financial ratios face several limitations. Accounting information, such as financial ratios, is based on past data, and it is usually determined one year before bankruptcy. That is, the bankruptcy prediction model based on financial ratios is considered a static model (Altman et al., 2010). A

time lag exists between the point of closing financial statements and the point of credit evaluation. In addition, financial ratios do not contain environmental factors such as external economic situations. Using only financial ratios may be insufficient in constructing a bankruptcy prediction model because they do not reflect the latest information, essentially reflecting past corporate internal accounting information.

Thus, adding qualitative information to the conventional bankruptcy prediction model is required to supplement accounting information. Some previous studies have attempted to use non-financial information other than internal accounting information such as types of business, firm age, and number of employees (Lee and Han, 1995; Grunert et al., 2005; Altman et al., 2010; Pervan and Kuvek, 2013), but these efforts still merely reflect internal non-financial information related to a firm due to lack of technologies for obtaining and analyzing qualitative information generated from external source. However, nowadays, vast amounts of data are obtainable on the web such as news, blog, and social network services (SNS). With this increasingly large amount of unstructured text data, big data analytics techniques, especially text mining, have been drawing much attention in academia and industry.

Qualitative information extracted from information sources on the web such as SNS postings, annual reports, and news can be complementary or alternative information for business prediction modeling in the several domains such as movie revenue prediction, political election prediction, customer churn prediction, and stock market prediction (Asur and Huberman, 2010; Coussement and Van den Poel, 2009; O'Connor et al., 2010; Schumaker et al., 2012; Tetlock et al., 2008). However, studies on the impact of qualitative information on the prediction model are still considered to be in the primary stage, restricted to limited applications such as stock prediction and movie revenue prediction applications. Thus, it is necessary to apply big data analytics techniques, such as text mining to various business prediction problems, including credit risk evaluation. Analytic methods are required for processing qualitative information represented in unstructured text form due to the complexity of managing and processing unstructured text data.

The purpose of this paper is to incorporate external qualitative information into the conventional bankruptcy prediction model to supplement limited accounting information and improve the predictive performance of conventional prediction models. Thus, the sentiment index is provided at the industry level by extracting from a large amount of text data to quantify the external economic atmosphere represented in the media. The proposed method involves keyword-based sentiment analysis using a domain-specific sentiment lexicon to extract sentiment from economic news articles. To create a specialized sentiment lexicon suitable for representing the construction business, the sentiment score of each term is assigned considering the relationship between the occurring

term and the actual situation with respect to the economic condition of the industry rather than the inherent semantics of the term.

The remainder of this paper is organized as follows. Section 2 provides previous studies on sentiment analysis applications. Section 3 provides the proposed method for incorporating qualitative information into the conventional bankruptcy prediction model. Section 4 describes the model development process, including research data and experimental designs. The experimental results and analysis are also described. Finally, Section 5 discusses the conclusions and future research issues.

2. Related work

Sentiment analysis originated is the computational study of sentiments represented in unstructured text. It is defined as the task of classifying positive or negative sentiments from unstructured text data. The challenges of detecting sentiments in text have been actively studied in recent years in various domains by using online reviews, SNS postings, and news articles. While the broader scope of the text mining technique focuses on the fact included in the text, sentiment analysis offers attitude. Sentiment analysis aims at classifying whether the text is subjective or objective (Wiebe et al., 2004; Yu and Hatzivassiloglou, 2003), whether it contains positive or negative sentiments (Pang et al., 2002; Turney, 2002), and whether the strength of polarity is weakly positive, mildly positive, or strongly positive (Pang and Lee, 2005; Wilson et al., 2004). We focus on classifying positive or negative sentiments, and two approaches are discussed for sentiment analysis: machine learning and lexicon-based. The studies on both the machine learning approach and lexicon-based approach are summarized in Table 1.

Machine learning approach adopts supervised learning since it performs the task of text classification by using a lot of training data labeled by sentiment. Labeled data is utilized to develop the sentiment classification model employing machine learning techniques (Boiy and Moens, 2009; Matsumoto et al., 2005; Melville et al., 2009; Pang et al., 2002; Sidorov et al., 2012; Ye et al., 2009).

Lexicon-based approach adopts unsupervised learning. It classifies the polarity of each word by using a set of sentiment words obtained by the existing sentiment lexicons or generating new ones. In the lexicon-based approach, two lexicon-based approaches are discussed for sentiment analysis: dictionary-based and corpus-based methods.

The dictionary-based approach has mainly exploited the general-purpose dictionary such as SentiWordNet based on the synonym and antonym relations of WordNet (Esuli and Sebastiani, 2005; Esuli and Sebastiani, 2006). In the corpus-based method, the pattern of word co-occurrence is considered, or new lexical resources are generated (Church and Hanks, 1990; Turney, 2002; Turney and Litman, 2003). Meanwhile, studies for

〈Table 1〉 Prior studies on sentiment analysis

Approach		Reference	Dataset	Method
Machine learning approach		Pang et al. (2002)	Movie reviews	SVM with unigram and feature presence
		Gamon (2004)	Customer feedback	SVM with feature selection based on log likelihood
		Matsumoto et al. (2005)	Movie reviews	SVM with bag-of-words feature and sub-pattern features
		Boiy and Moens (2009)	Blog posting, Customer reviews, News forums	SVM, Naïve Bayes with unigram and language-specific features
		Melville et al. (2009)	Blog posting	Naïve Bayes with lexical knowledge
		Ye et al. (2009)	Travel destination reviews	SVM with N-gram model
		Sidorov et al. (2012)	Twitter posting	SVM with unigram
Lexicon-based approach	Dictionary-based method	Kamps et al. (2004)	Product review	WordNet
		Hu and Liu (2004)	Product review	WordNet
		Kim and Hovy (2004)	100 sentences from DUC 2001 corpus	WordNet
		Esuli and Sebastiani (2005, 2006)	WordNet	SentiWordNet
	Corpus-based method	Turney (2002)	Automobile reviews, Movie reviews	Point-wise mutual information (PMI)
		Turney and Litman (2003)	Automobile reviews, Movie reviews	Point-wise mutual information (PMI) and latent semantic analysis (LSA)
		Ding et al. (2008)	Product reviews	Domain-specific sentiment lexicons
		Du et al. (2010)	Hotel reviews, Electronics reviews, Stock reviews	Integrating the cross-domain knowledge and within-domain knowledge
		Song and Lee (2011)	Product reviews	Domain-specific sentiment lexicons
		Salah et al. (2013)	Political debate data	Domain-specific sentiment lexicons
		Yu et al. (2013)	News articles	Domain-specific sentiment lexicons
		Kim and Kim (2014)	Movie reviews	Domain-specific sentiment lexicons
		Jeong et al. (2015)	News articles	Domain-specific sentiment lexicons

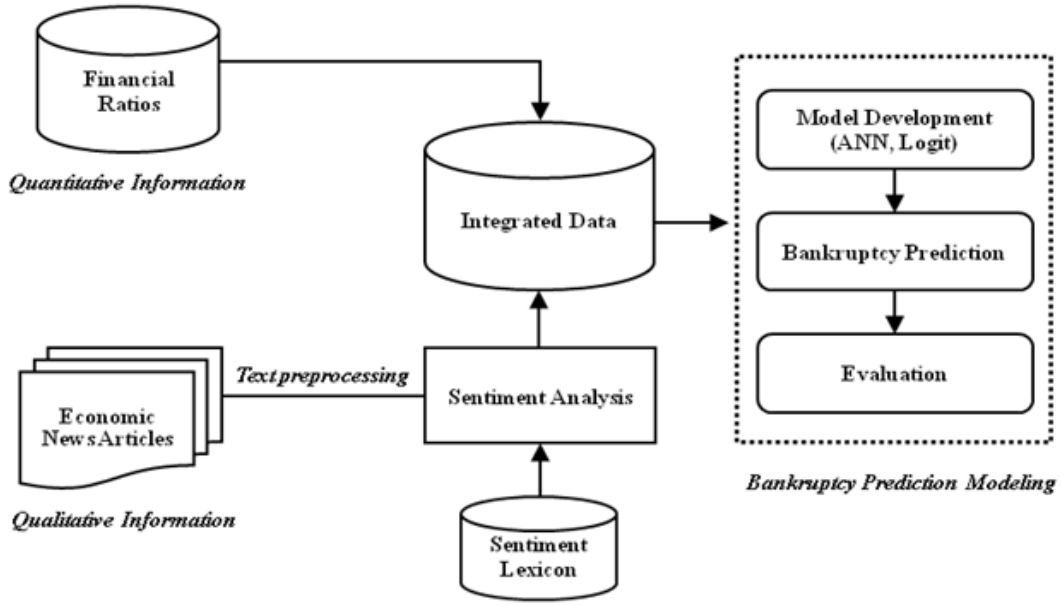
applying general-purpose lexicons to new lexicons generated by the domain corpus have been performed (Du et al., 2010; Salah et al., 2013). Although the SentiWordNet has widely used in sentiment analysis studies, it is difficult to correctly identify sentiment focused on Korean contexts since it is constructed based on English. There are a few studies that perform sentiment analysis based on a domain-specific sentiment lexicon based on Korean language (Song and Lee,

2011; Jeong et al., 2015; Kim and Kim, 2014; Yu et al., 2013). Conventional sentiment analysis techniques did not consider the flexibility of polarity, which could be changed by the situation, and the polarity of word was determined by not inherent semantics but its meaning in a particular context.

Most of the studies on sentiment analysis have used explicit opinions represented in a subjective text such as online reviews and SNS postings. Relatively few studies have done sentiment analysis by utilizing implicit opinions represented in objective texts such as news articles (Kim and Hovy, 2006). Although the news articles represent an objective statement, objective text includes sentiments reflected by desirable and undesirable facts (Zhang and Liu, 2011). Thus, detecting sentiment based on domain context rather than considering inherent meaning of words is necessary because implicit opinions are rather difficult to identify the sentiment compared with explicit opinions.

3. Proposed model

The proposed unified framework for bankruptcy prediction modeling employs quantitative and qualitative information based on big data analytics as shown in <Figure 1>. The proposed model consists of three tasks: sentiment lexicon generation from economic news, sentiment analysis using a domain-specific sentiment lexicon, and bankruptcy prediction modeling using both



〈Figure 1〉 Unified framework for bankruptcy prediction modeling using financial and sentiment variables

financial ratios and a sentiment variable.

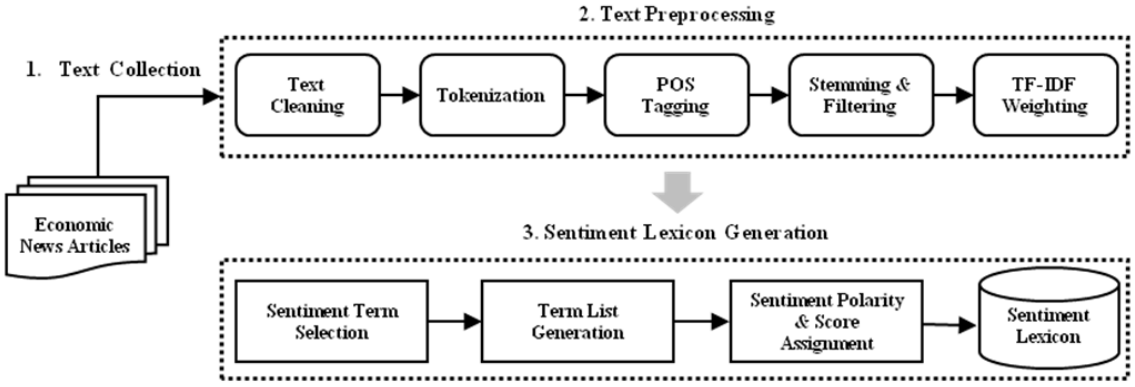
Bankruptcy prediction models are constructed categorizing three parts. First, the basic model uses financial ratios, and logit and ANN for constructing bankruptcy prediction models. Second, the CBI model uses financial ratios and a construction business index to reflect the economic situation in the construction industry. Third, the proposed model inserts a sentiment variable extracted from economic news into the basic model using only financial ratios. It utilizes financial ratios and a sentiment variable representing industry economic index as an auxiliary input variable simultaneously for constructing the ANN model. The sentiment variable is derived from contents of news articles after the point of closing financial statements date, and finishes news observation before the credit evaluation date by considering that there exists a time lag between the point of closing financial statements and the point of credit evaluation.

3.1 Generating a domain-specific sentiment lexicon

The first task of the proposed method is to generate sentiment lexicon utilizing economic news articles. This is done here by adopting a method for generating a new domain-specific lexicon to extract useful sentiment score from economic news articles. When using a general-purpose sentiment lexicon applied to all domains, it is difficult to apply the semantic orientation of terms that could be used differently according to the domain. <Figure 2> shows the process of generating a domain-specific sentiment lexicon for sentiment analysis.

3.1.1. Text preprocessing

Text mining, first mentioned by Feldman and Dagan (1995), refers to an approach to capture useful information from data by exploring interesting patterns and trend analysis in the unstructured text data. The purpose of text mining



〈Figure 2〉 Process of generating a sentiment lexicon

is to transform unstructured text data to structured numerical data to extract meaningful knowledge. The architecture of text mining is composed of four tasks: document collection, preprocessing, mining, and visualization. Text preprocessing tasks are preceded prior to extract knowledge through mining such as text clustering and text classification. These tasks convert natural language from raw document into a structured form before applying text feature selection and finding patterns and trends; thus creating a new document collection represented by words (Feldman and Sanger, 2007).

In this study, economic news articles related to the construction business from two major economic newspapers in Korea are crawled. Extracting key terms in a large number of news articles requires text preprocessing for transforming the format from unstructured into structured data. Text preprocessing consists of the following basic process: cleaning, tokenization, POS tagging, stemming, term extraction, term filtering, and term weighting.

We first remove numbers, special characters, and punctuation from the news collection by performing a cleaning task, and then tokenize the news articles into individual terms. POS tagging assigns individual words as their syntactic category. In this study, we retrieve terms with the POS of adjectives, verb, and noun. In particular, the POS of adjectives is critical information in finding both subjectivity and polarity in text data.

Next, stemming considers different terms having the same root as the same terms. Then, we remove terms with low frequency or stopwords through term filtering. Stopwords include non-informative terms which are not proper for representing the domain and the definite article such as ‘the’ or ‘a’.

Term weighting identifies important terms compared with other terms using TF-IDF (Term Frequency-Inverse Document Frequency) index which is a widely used term weighting method (Salton and Buckley, 1988), rather than using only term frequency. TF (Term frequency) denotes a value that indicates the frequency of a particular term appearing within a document. Although the high TF represents that the term is important in a document, it means that frequently occurring terms has low importance as the common terms. Thus, IDF (Inverse Document frequency) proposed by Sparck Jones (1972) is incorporated to reflect how commonly a specific term occurs within a document collection. It is computed by dividing the the total number of all documents by the number of documents including the term, and then taking a logarithm of IDF. TF-IDF is calculated by multiplying TF by IDF values. The formula of TF-IDF has two parts, TF and IDF, to compute term weight as follows.

$$w_{ij} = TF_{ij} IDF_i \quad (1)$$

$$TF_{ij} = log_2(f_{ij}+1) \quad (2)$$

$$IDF_i = log_2(n / df_i) \quad (3)$$

where

TF_{ij} : the number of occurrences of $term_i$ in $document_j$

IDF_i : inverse document frequency of $term_i$

f_{ij} : frequency of $term_i$ in $document_j$
n : the total number of all documents
df_i : the number of documents containing $term_i$

The unstructured text is converted into a structured term-document matrix composed of rows representing terms and columns representing documents through the process of text preprocessing.

3.1.2. Sentiment lexicon generation

Initial terms are discovered from the news collection after the text preprocessing tasks are completed. The secondary terms are generated by considering two types of characteristics of a term, semantic and statistical aspects. Terms whose minimum TF-IDF weight is greater than 1 and terms for which semantic orientation is not appropriate are removed. Consequently, we generate the final term list of the sentiment lexicon, categorizing three types of terms: positive, negative, and neutral terms.

Sentiment indicates commonly used terms to express positive or negative feelings, e.g., ‘good,’ ‘great,’ ‘bad,’ and ‘poor.’ The final sentiment of a document is determined based on the occurrence frequency of positive and negative sentiment terms. Although it is a simple and commonly adopted method, it has limitations such that terms related to domain context cannot be handled (Ding et al., 2008). Thus, sentiment score is also assigned for domain terms even if it is not restricted to general sentiment terms such as ‘happy’ and ‘sad.’ The sentiment lexicon is generated by assigning the polarity and the sentiment score of each term in the term list. To create a specialized sentiment lexicon, it is advisable to assign the sentiment score of each term depending on the relationship between the term and the domain context rather than its inherent semantics. The proposed approach uses both a data-driven and human generated method to make a term list considering various POSs such as nouns, verbs, adverbs, and adjectives.

The sentiment score is computed based on the occurrence frequency of each term in the economic situation associated with the industry. The range of the sentiment score is from -1 to 1, with scores closer to the former indicating negative sentiment and scores closer go the latter positive. The sentiment score of each term in the final term list was calculated as follows:

$$Score(term_i) = \frac{good(term_i) - bad(term_i)}{N} \quad (4)$$

where
N : the number of news articles occurrences of $term_i$
$good(term_i)$: the number of news articles occurrences of $term_i$ in a good economic situation in the construction industry
$bad(term_i)$: the number of news articels occurrences of $term_i$ in a bad economic situation in the construction industry

If the sentiment score is higher than 0, the sentiment polarity of $term_i$ was identified as positive; If the sentiment score is lower than 0, it was identified as negative. Terms for which the sentiment score is zero are considered as neutral terms, and those are excluded from the term list in sentiment analysis.

The economic situation of the construction industry is detected by the economic index for the construction industry, the Construction Business Survey Index (CBSI) developed by the Construction Economy Research Institute of Korea. A good economic situation in the construction industry is represented when the index related to construction is ascendant; otherwise, a bad economic situation in the construction industry is represented when the index is moving downward. This index is used to quantify the opinion for the future economic trends from entrepreneurs, and used for a short-term economic prediction index. The proposed model uses a sentiment index developed in this study based on news articles related to construction business and CBSI as auxiliary data. CBSI is appropriate for use in the development of a sentiment index for the construction industry since it is the subjective and psychological factors are reflected, unlike with other indices for the construction industry such as construct investment and construction contracts.

The proposed sentiment index can identify economic phenomenon and problems pertaining to the economy of the construction industry that are difficult to capture using the previous economic index itself through big data analytics based on economic news articles. It is possible to investigate causes of the economic phenomenon in the construction industry by semi-automatically extracting from a large amount of text data.

3.2 Sentiment analysis using a domain-specific sentiment lexicon

Unlike user-generated content such as online review and SNS postings, the news article typically represents more structured and implicit sentiment. Generating a sentiment lexicon fitted to terms occurring in the news articles is required for extracting sentiment from news expressed by more subtle opinion. Most studies on sentiment analysis follow a common approach that first classifies positive and negative terms, and then sums or averages values of sentiment scores. In this study, we adopt a scoring method at the keyword level. Terms used in the term-document matrix for sentiment analysis are represented by unigram and feature presence. Using features with unigram and feature presence has been reported to be superior to employing complex forms of features in dealing with the problem of sentiment classification (Kim and Hovy, 2004; Pang et al., 2002; Yu and Hatzivassiloglou, 2003).

For transforming news articles composed of many terms into sentiment, a sentiment lexicon is generated including a set of terms with positive or negative orientations and their score. We confirm a term-document matrix to examine the existence of sentiment terms within documents. The sentiment term-document matrix weighted by

sentiment score requires capturing sentiment of each news article. The matrix is composed by multiplying a term-document matrix and sentiment score vector from the sentiment lexicon.

The overall sentiment score of single news is computed by summing up or averaging scores of sentiment words within a news article. The overall sentiment score of all referenced news during the corresponding period for each firm is computed by several sentiment measures: average sentiment score (S1) of total news occurring during the period of news observation corresponding to each firm; the difference between percentage of positive articles and percentage of negative articles (S2); the ratio of positive to negative articles (P-ratio); and the ratio of negative to positive articles (N-ratio).

S1 and S2 indicate overall sentiment degree of total news. P-ratio reflects the assumption that a firm that has more positive articles than negative articles are likely to be healthy. N-ratio reflects the assumption that a firm that has more negative than positive articles is likely to be bankrupt. Each sentiment measure is computed as follows:

$$S1 = \frac{\sum_{j=1}^{t} Sum(news_j)}{t} \quad (5)$$

$$S2 = \frac{Count(Pos) - Count(Neg)}{t} \quad (6)$$

$$P-ratio = \frac{Count(Pos)}{Count(Neg)} \quad (7)$$

$$N-ratio = \frac{Count(Neg)}{Count(Pos)} \quad (8)$$

where

$Sum(news_j)$: sum of sentiment score of jth single news

t : the number of news articles occurring during the specified period of news observation after the point of the closing financial statements

$Count(Pos)$: the number of news articles with positive sentiment

$Count(Neg)$: the number of news articles with negative sentiment

If the average sentiment score of single news article is higher than 0, the sentiment polarity of single news was identified as positive; If the average sentiment score is lower than 0, it was identified as negative. News articles for which the average sentiment score is zero are considered as neutral news articles, and those are excluded from the reference news articles in sentiment analysis.

Sentiment indices are developed here based on total news articles occurring in the corresponding period of the news observation for each firm by using sentiment score of single news articles. The period of news observation is set to reflect the qualitative information between the point of closing financial statements and the point of credit evaluation. We select a prediction model with a time lag showing high classification accuracy by applying various time lags from 1 to 5 months from the point of closing financial statements in extracting sentiment from news articles.

4. Model development

4.1 Research data and experiments

The research data consisted of quantitative and qualitative information. As quantitative information, financial data was derived from a financial statement of 916 Korean small and medium-sized construction firms, 458 firms for bankruptcy and the other 458 firms for non-bankruptcy from 2008 to 2012. We randomly selected 458 non-bankrupt firms from all solvent firms. The data set was split into two subsets; 80% of the data is used for a training set and 20% for holdout set. The training set was used for model development and divided into training set (60%) and test set (20%). The test set was used to avoid over-fitting of the ANN model in order to generalize well on new data. The holdout set was used to test the validity of the model by using the data that is not used to construct the model.

Two steps were implemented to select input fianacial variables. In the first stage, the number of financial ratios was reduced from 61 to 16, as selected by univariate test and the opinions of experts. In the second stage, the five input variables were chosen by using a stepwise method to reduce the dimensionality. The selected variables for this study are shown in Table 2.

As qualitative information, news articles data were collected from M and H major business presses published in Korea, and retrieved 81,318 daily news articles of economy section including keyword ‘construction,’ and duplicate news articles were eliminated. This study was aimed at analyzing the overall aspects of economic situation in the construction industry, so more sophisticated terms, including ‘construction business’ or ‘construction industry business,’ were chosen as additional search keywords. We refined 81,318 news articles related to construction into 2,274 news articles through constructing a domain corpus directly related to the construct business.

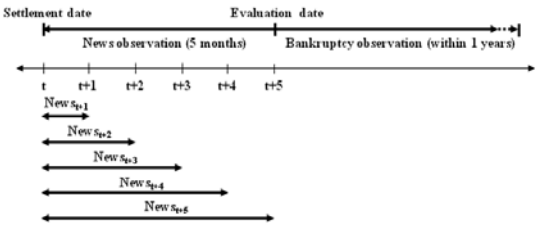
〈Table 2〉 Definition of financial variables

Variable	Definition
X1	Financial expenses to sales
X2	Total borrowings and bonds payable to total assets
X3	Capital adequacy ratio
X4	Productivity of capital, gross value-added to total assets
X5	Turnover ratio of working capital

The news data was split into a sentiment lexicon generation set and application set. The lexicon generation set of news data was used for text preprocessing such as stemming and TF-IDF weighting and calculation of sentiment score and constructing a sentiment lexicon. The application data was used to test the adequacy on the list of sentiment terms constructed in the sentiment lexicon generation set.

The collection of news articles was refined and the overall sentiment calculated using average sentiment score (S1) and the difference between percentage of positive articles and percentage of negative articles (S2), positive sentiment (P-ratio),

and negative sentiment (N-ratio). The period of bankruptcy observation was set within in 1 year after the credit evaluation date. There was a time lag between the point of closing financial statements and the point of credit evaluation. Thus, we observed news articles after the settlement date and finished news observation at the credit evaluation date as shown in <Figure 3>. We partition news data with various periods of news observation to test the appropriateness of the time lag, and reflect news information during the period of from 1-5 months (i.e., from time t+1 to t+5) after settlement date at time t.



〈Figure 3〉 Period of news observation

The multilayer perceptron (MLP) network trained by the back-propagation (BP) algorithm was selected. The sigmoid transfer function was used in the hidden and output nodes. We set the number of hidden nodes as the number of input variables in all developed models. One output node was used for a binary classification. Bankruptcy was defined in terms of outputs, and the range of outputs is [0, 1]. The learning rate was set to 0.1 and momentum was set to 0.1. The learning epochs were set to 1,000 for all experiments. The ANN models were developed by using the software Neuroshell 2.

4.2 Result and Analysis

To examine the effectiveness of the proposed model, which incorporates qualitative information in the context of the bankruptcy prediction problem, text mining techniques among big data analytics were adopted to transform qualitative into quantitative information. To classify the polarity of sentiment from economic news articles, the keyword-based sentiment analysis was used to generate the domain-specific sentiment lexicon.

We got an initial 20,000 terms by going through text parsing from news collection. The secondary term list included 2,935 terms by filtering process that removes terms whose minimum TF-IDF is greater than 1 to improve statistical quality and terms for which the number of documents containing the term falls short of minimum 13 documents (about 1% of the lexicon generation set size). Finally, 367 terms were obtained by removing terms that semantic orientation was not appropriate.

367 terms were classified into three categories; 72 positive terms, 130 negative terms, and 146 neutral terms. The number of terms with negative sentiments was greater than that of terms with positive sentiment in the extracted terms from news articles about construction business, probably because the perspective on the construction market is often negative in the media. The sentiment lexicon contained the polarity of sentiments and their scores, which indicate the sentiment intensity

〈Table 3〉 Example of semantic orientation classification result

Sentiment terms	Number of good news articles	Number of bad news articles	Score	Polarity	Evaluation
Unsold	111	154	-0.162	Negative	○
Drop	91	138	-0.238	Negative	○
Stable	40	52	-0.130	Negative	×
Benefit	27	42	-0.217	Negative	×
Opportunity	39	27	0.182	Positive	○
Revival	14	6	0.400	Positive	○
Adopt	25	25	0.000	Neutral	×

of a term. The sentiment of terms was determined by the number of news articles including each term when the economic situation was good or bad as shown in <Table 3>.

For semantic orientation of each term when the sentiment score was higher than 0, the sentiment polarity was positive; when the sentiment score was lower than 0, it was negative. If the sentiment score was 0, the term was neutral, and excluded; the resulted in 221 words for sentiment analysis.

<Table 4> presents an example of the generated sentiment terms in the sentiment lexicon. The range of the sentiment score was from -1 to 1. The sentiment score of each term was assigned using a

〈Table 4〉 Example of a sentiment lexicon for the construction business

No.	Positive			Negative		
	Term	POS	Score	Term	POS	Score
1	Heyday	Noun	0.714	Warning	Noun	-0.846
2	Boast	Verb	0.714	Bankruptcy	Noun	-0.750
3	Certification	Noun	0.692	Criticism	Noun	-0.500
4	Stability	Noun	0.692	Slowdown	Noun	-0.448
5	Revival	Noun	0.400	Descent	Noun	-0.444
6	Concentration	Noun	0.385	Falter	Verb	-0.444
7	Added value	Noun	0.364	Limitation	Noun	-0.400
8	Influence	Noun	0.333	Sigh	Noun	-0.385
9	Survival	Noun	0.333	Dull	Adj	-0.368
10	Rush into	Verb	0.290	Credit crunch	Noun	-0.360
11	Efficiency	Noun	0.250	Inflation	Noun	-0.360
12	Attractive	Adj	0.238	Worry	Noun	-0.333
13	Leap	Noun	0.231	Urgent	Adj	-0.333
14	Green growth	Noun	0.231	Anxious	Adj	-0.333
15	Powerful	Adj	0.200	Pressure	Verb	-0.333

specialized sentiment lexicon of the construction business based on the relationship between the term and economic situation. The sentiment score of each term was determined with respect to the domain rather than the inherent semantics of each term. The polarity and the score of each term were based on the economic situation.

We defined sentiment variables based on external information such as economic news articles and CBSI index. The period of news observation was set to reflect the qualitative information between the point of closing financial statements and the point of credit evaluation. The sentiment score from economic news and financial ratios were used as the input variables for ANN. Independent t-test was conducted to examine whether the sentiment variable defined in this study is significant indicator in discriminating bankruptcy firms.

The result of univariate test showed that news sentiment variables (S1, S2, P-ration, and N-ratio) were significant at a 1% significance level. Thus, news sentiment score was useful in explaining the mean difference between bankruptcy and non-bankruptcy firms. To investigate the difference in classification accuracy according to the period of reflecting sentiment extracted from economic news in the context of bankruptcy prediction problems, the experiment was performed with respect to various time lags between settlement date and credit evaluation date.

A basic bankruptcy prediction model using only financial ratios, constructed using logit analysis, correctly classified 67.94% for the training set and 66.12% for the holdout set. The number of hidden nodes was determined by experiments. A BPN model with six hidden nodes showed the best classification accuracy. A basic bankruptcy prediction model using only financial ratios by BPN correctly classified 73.09% for the training set, 71.04% for the test set, and 69.95% for the holdout set.

Incorporating sentiment variables into conventional bankruptcy prediction models based on financial ratios was constructed by BPN models with seven hidden nodes. The comparison of the results according to the period of reflecting news of the bankruptcy prediction model suggested for this study is shown in <Tables 5>.

〈Table 5〉 Classification accuracy of models

Model	Period	Training	Test	Holdout
S1	$News_{t+1}$	77.64%	74.32%	74.32%
	$News_{t+2}$	79.09%	76.50%	77.60%
	$News_{t+3}$	73.27%	70.49%	66.67%
	$News_{t+4}$	74.36%	70.49%	68.31%
	$News_{t+5}$	74.00%	70.49%	68.31%
S2	$News_{t+1}$	76.36%	73.22%	70.49%
	$News_{t+2}$	75.45%	71.58%	74.86%
	$News_{t+3}$	75.64%	71.04%	70.49%
	$News_{t+4}$	76.91%	72.13%	70.49%
	$News_{t+5}$	75.82%	72.13%	66.12%
P-ratio	$News_{t+1}$	76.36%	73.22%	70.49%
	$News_{t+2}$	75.45%	71.58%	74.86%
	$News_{t+3}$	75.64%	71.04%	70.49%
	$News_{t+4}$	76.91%	72.13%	70.49%
	$News_{t+5}$	75.82%	72.13%	66.12%
N-ratio	$News_{t+1}$	77.82%	71.58%	68.31%
	$News_{t+2}$	79.45%	74.86%	79.23%
	$News_{t+3}$	77.45%	72.13%	71.58%
	$News_{t+4}$	78.00%	71.58%	71.58%
	$News_{t+5}$	75.64%	69.40%	73.22%

The results of classification accuracy on the holdout set according to the period of reflecting sentiment are summarized in <Table 6>. A BPN model incorporating the suggested four sentiment variables showed the best classification accuracy on the holdout set in time t+2. Even though the amount of information reflected for sentiment extraction increases, it was found that classification accuracy of the model tends to decrease, as shown in <Figure 4>. That is, it is interpreted that the predictive performance of the model was improved the smaller the interval of reflecting qualitative information between the point of closing financial statements and the point of credit evaluation. Consequently, time t+2 was selected as the period of reflecting news sentiment.

〈Table 6〉 Classification accuracy on the period of reflecting sentiment information

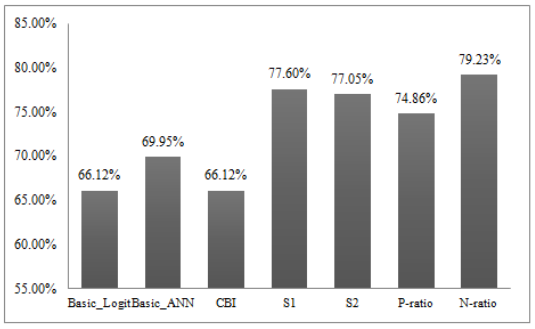
Model	$News_{t+1}$	$News_{t+2}$	$News_{t+3}$	$News_{t+4}$	$News_{t+5}$
S1	74.32%	77.60%	66.67%	68.31%	68.31%
S2	74.32%	77.05%	67.21%	67.21%	68.31%
P-ratio	70.49%	74.86%	70.49%	70.49%	66.12%
N-ratio	68.31%	79.23%	71.58%	71.58%	73.22%

〈Figure 4〉 Comparison of classification accuracy on the period of reflecting sentiment information

<Table 7> and <Figure 5> show the comparison of the predictive performance on the bankruptcy prediction model suggested for this study. The experimental results of the proposed models were compared with those of the conventional bankruptcy prediction models using logit analysis and BPN based on financial ratios as quantitative information. Overall, the classification accuracies of S1, S2, P-ratio, and N-ratio models proposed for this study were higher than those of logit and BPN using financial ratios, as well as a model using a construction business index.

〈Table 7〉 Comparison of classification accuracy

Model	Training	Test	Holdout
Basic_Logit	67.94%		66.12%
Basic_ANN	73.09%	71.04%	69.95%
CBI	61.09%	60.11%	66.12%
S1	79.09%	76.50%	77.60%
S2	78.91%	74.32%	77.05%
P-ratio	75.45%	71.58%	74.86%
N-ratio	79.45%	74.86%	79.23%



〈Figure 5〉 Classification performance of the holdout set

It is concluded that incorporating qualitative information based on sentiment analysis into the bankruptcy prediction model based on accounting information enhances the classification performance. In particular, the N-ratio model combining negative sentiment showed the best classification accuracy among the proposed models. Adding the sentiment variable as qualitative information in a bankruptcy prediction model plays an important role in improving the predictive performance, supplementing the limitation of using only accounting information.

We conducted the McNemar test to confirm whether the performance of the proposed model was significantly higher than those of the other basic model. The McNemar test is a nonparametric test used on paired nominal data. It was attempted to determine whether there were significant changes of classification accuracy rate (the number of correct classifications by the number of whole Holdout samples). Table 8 shows McNemar values for classification accuracy between models.

The result of the McNemar test proved that the proposed models using sentiment variables have higher predictive performance than not only the conventional bankruptcy prediction model but also the model including the construction business index under the same learning condition of ANN. Thus, it was concluded that the sentiment indicators contributed to discriminating bankruptcy firms. The predictive performance among the proposed models using sentiment variables for this study represented similar levels in terms of classification accuracy. However, the predictive performance of N-ratio model (79.23%) outperformed that of P-ratio model (74.86%). Negative sentiment had a greater impact on predicting corporate bankruptcy compared to positive sentiment on media.

5. Conclusion

This paper examined the benefit of incorporating sentiment variables as qualitative information to a conventional bankruptcy

〈Table 8〉 McNemar values for the comparison of performance between models

(Significance level)

	Basic_ANN	CBI	S1	S2	P-ratio	N-ratio
Basic_ Logit	0.248	1.000	0.001***	0.004***	0.014**	0.000***
Basic_ ANN	-	0.483	0.007***	0.011**	0.078*	0.001***
CBI	-	-	0.008***	0.012**	0.044**	0.001***
S1	-	-	-	1.000	0.227	0.453
S2	-	-	-	-	0.344	0.219
P-ratio	-	-	-	-	-	0.039**

*** Significant at 1%, ** Significant at 5%, * Significant at 10%

prediction model based on financial ratios as quantitative information. In this study, we constructed a domain-specific sentiment lexicon from economic news related to the construction business for deriving sentiment. Experimental results support the following meaningful findings. First, a unified framework for using sentiment variable and financial ratios significantly enhances the predictive performance of conventional bankruptcy prediction models. Second, the final model using news articles during two months to extract sentiment score was selected through experiments of various time lags. The predictive performance of the model was improved the smaller the interval between financial information of the point of closing financial statements and news of the point of closing. Third, the sentiment extracted from economic news articles had a significant impact on corporate bankruptcy. In particular, negative sentiment was better able to predict corporate bankruptcy. This indicates that corporate bankruptcy of construction firms is sensitive to poor economic conditions. However, it is difficult to apply the sentiment variables which are derived from this research to develop the bankruptcy prediction model, because they are sentiment indices which reflect information on the construction industry, not individual firms.

This study has the following limitations that require future study. First, this study divided a corpus into two datasets for sentiment analysis: dataset consisting of a sentiment lexicon generation set and one for applying sentiment score of each term extracted from the sentiment lexicon generation set. It is dependent on the economic conditions of the time at which the text was generated. Thus, the sentiment lexicon should be continually updated since the terms are likely to be destroyed over time. Moreover, it is considered to analyze except for news articles included in special events that may cause the contemporary bias, or to acquire corpus of a longer period.

Second, it is necessary to construct a corpus by manually labeling each item therein with the unknown target variable when applying sentiment classification. News articles are generally written in neutral tone, and are, therefore, hard to assign actual sentiments such as positive and negative. In this study, we utilized the movement direction of the construction business index as the target variable of the corpus for sentiment analysis. However, verification of whether classifying polarity for each news articles is appropriate or not is difficult because not all news articles are labeled by actual sentiments. Obtaining a corpus assigned by polarity through expert opinions of the construction economic sector should be considered.

Lastly, the performance of sentiment analysis is dependent on the lexicon resource. A sentiment lexicon that contains the appropriate amount of sentiment terms is crucial to improving the classification accuracy of the sentiment analysis result. Thus, we need to consider the optimizing issue in sentiment analysis about parameters such as how the number of sentiment terms, term selection, and threshold for polarity detection can be optimized for improving accuracy of sentiment classification.

References

Altman, E. I., "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy," *The journal of finance*, Vol.23, No.4(1968), 589~609.

Altman, E. I., Sabato, G., and N. Wilson, "The value of non-financial information in small and medium-sized enterprise risk management," *Journal of Credit Risk*, Vol.2, No.6(2010), 95~127.

Asur, S. and B. A. Huberman, "Predicting the future with social media," *Proceedings of 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, Vol.1, (2010), 492~499.

Boiy, E. and M. F. Moens, "A machine learning approach to sentiment analysis in multilingual Web texts," *Information retrieval*, Vol.12, No.5(2009), 526~558.

Church, K. W. and P. Hanks, "Word association norms, mutual information, and lexicography," Computational linguistics, Vol.16, No.1(1990), 22~29.

Coussement, K. and D. Van den Poel, "Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers," *Expert Systems with Applications*, Vol.36, No.3(2009), 6127~6134.

Ding, X., Liu, B., and P. S. Yu, A holistic lexicon-based approach to opinion mining. *Proceedings of the 2008 International Conference on Web Search and Data Mining*, ACM, 2008, 231~240.

Du, W., Tan, S., Cheng, X., and X. Yun, "Adapting information bottleneck method for automatic construction of domain-oriented sentiment lexicon," *Proceedings of the third ACM international conference on Web search and data mining*, ACM, 2010, 111~120.

Esuli, A. and F. Sebastiani, "Determining the semantic orientation of terms through gloss classification," *Proceedings of the 14th ACM international conference on Information and knowledge management*, 2005, 617~624.

Esuli, A. and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining," *Proceedings of LREC*, Vol.6(2006), 417~422.

Feldman, R. and I. Dagan, "Knowledge Discovery in Textual Databases (KDT)," *KDD*, Vol.95 (1995), 112~117.

Feldman, R. and J. Sanger, *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*, Cambridge University Press, 2007.

Fletcher, D. and E. Goss, "Forecasting with neural networks: an application using bankruptcy data," *Information & Management*, Vol.24, No.3(1993), 159~167.

Grunert, J., Norden, L., and M. Weber, "The role of non-financial factors in internal credit ratings," *Journal of Banking & Finance*, Vol.29, No.2(2005), 509~531.

Hamer, M. M., "Failure prediction: sensitivity of classification accuracy to alternative statistical methods and variable sets," *Journal of Accounting and Public Policy*, Vol, 2, No.4 (1984), 289~307.

Jeong, J. S., D. S. Kim, and J. W. Kim, "Influence analysis of Internet buzz to corporate performance : Individual stock price

prediction using sentiment analysis of online news," *Journal of Intelligence and Information Systems*, Vol.21, No.4(2015), 37~51.

Kim, S. M. and E. Hovy, "Determining the sentiment of opinions," *Proceedings of the 20th international conference on Computational Linguistics*, Association for Computational Linguistics, (2004), 1367~1373.

Kim, S. M. and E. Hovy, "Extracting opinions, opinion holders, and topics expressed in online news media text," *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, Association for Computational Linguistics, 2006, 1~8.

Kim, S. and N. Kim, "A Study on the Effect of Using Sentiment Lexicon in Opinion Classification," *Journal of Intelligence and Information Systems*, Vol.20, No.1(2014), 133~148.

Lee, J. S. and J. H. Han, "Usability Test of Non-Financial Information in Bankruptcy Prediction using Artificial Neural Network-The Case of Small and Medium-Sized Firms," *Journal of Intelligence and Information Systems*, Vol.1, No.1(1995), 123~134.

Leshno, M. and Y. Spector, "Neural network prediction analysis: The bankruptcy case," *Neurocomputing*, Vol.10, No.2(1996), 125~147.

Matsumoto, S., Takamura, H., and M. Okumura, "Sentiment classification using word sub-sequences and dependency sub-trees," *Proceedings of the 9th Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining*, Springer-Verlag, 2005, 301~311.

Melville, P., Gryc, W., and R. D. Lawrence, "Sentiment analysis of blogs by combining lexical knowledge with text classification," *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009, 1275~1284.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., and N. A. Smith, "From tweets to polls: Linking text sentiment to public opinion time series," *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, Vol.11(2010), 122~129.

Odom, M. D. and R. Sharda, "A neural network model for bankruptcy prediction," *Proceedings of IJCNN International Joint Conference on Neural Networks*, IEEE, 1990, 163~168.

Ohlson, J. A., "Financial ratios and the probabilistic prediction of bankruptcy," *Journal of accounting research*, Vol.18, No.1(1980), 109~131.

Pang, B. and L. Lee, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales," *Proceedings of the Association for Computational Linguistics (ACL)*, 2005, 115~124.

Pang, B., Lee, L., and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," *Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, Vol.10(2002), 79~86.

Pervan, I. and T. Kuvek, "The Relative Importance of Financial Ratios and Nonfinancial Variables in Predicting of Insolvency," *Croatian Operational Research Review*, Vol.4, No.1(2013), 187~197.

Salah, Z., Coenen, F., and D. Grossi, "Generating

Domain-Specific Sentiment Lexicons for Opinion Mining," *Advanced Data Mining and Applications*, Springer Berlin Heidelberg, 2013, 13~24.

Salton, G. and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing and Management*, Vol.23, No.5(1988), 513~523.

Schumaker, R. P., Zhang, Y., Huang, C. N., and H. Chen, "Evaluating sentiment in financial news articles," *Decision Support Systems*, Vol.53, No.3(2012), 458~464.

Shaw, M. J. and J. A. Gentry, "Inductive learning for risk classification," *IEEE Expert*, Vol.5, No.1(1990), 47~53.

Shin, K.-s., Lee, T. S., and H,-j. Kim, "An application of support vector machines in bankruptcy prediction model," *Expert Systems with Applications*, Vol.28, No.1(2005), 127~135.

Sidorov, G. et al., "Empirical study of machine learning based approach for opinion mining in tweets," *Proceedings of the 11th Mexican international conference on Advances in Artificial Intelligence*, Vol. Part Ⅰ, 2012, 1~14.

Song, J. and S. Lee, "Automatic Construction of Positive/Negative Feature-Predicate Dictionary for Polarity Classification of Product Reviews," *Journal of KIISE: Software and Applications*, Vol.38, No.3(2011), 157~168.

Sparck Jones, K., "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, Vol.28, No.1(1972), 11~21.

Tam, K. Y. and M. Y. Kiang, "Managerial applications of neural networks: the case of bank failure predictions," *Management science*, Vol.38, No.7(1992), 926~947.

Tetlock, P. C., "Saar-Tsechansky, M., and S. Macskassy, "More than words: Quantifying language to measure firms' fundamentals," *The journal of finance*, Vol.63, No.3(2008), 1437~1467.

Turney, P. D., "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews," *Proceedings of the 40th annual meeting on association for computational linguistics*, 2002, 417~424.

Turney, P. D. and M. L. Littman, "Measuring praise and criticism: Inference of semantic orientation from association," *ACM Transactions on Information Systems (TOIS)*, Vol.21, No.4(2003), 315~346.

Wiebe, J., Wilson, T., Bruce, R., Bell, M., and M. Martin, "Learning subjective language," *Computational linguistics*, Vol.30, No.3(2004), 277~308.

Wilson, T., Janyce W., and R. Hwa, "Just how mad are you? Finding strong and weak opinion clauses," *Proceedings of National Conference on Artificial Intelligence (AAAI-2004)*, 2004, 761~767.

Ye, Q., Zhang, Z., and R. Law, "Sentiment classification of online reviews to travel destinations by supervised machine learning approaches," *Expert Systems with Applications*, Vol.36, No.3(2009), 6527~6535.

Yu, E., Kim, Y., Kim, N., and S. R. Jung, "Predicting the Direction of the Stock Index by Using a Domain-Specific Sentiment Dictionary," *Journal of Intelligence and Information Systems*, Vol.19, No.10(2013), 95~110.

Yu, H. and V. Hatzivassiloglou, "Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences," *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, 2003, 129~136.

Zhang, L. and B. Liu, "Identifying noun product features that imply opinions," *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, Vol.2, (2011), 575~580.

국문요약

빅데이터 기반의 정성 정보를 활용한 부도 예측 모형 구축*

조남옥** · 신경식***

대부분의 부도 예측에 관한 연구는 재무 변수를 중심으로 통계적 방법 또는 인공지능 기법을 적용하여 부도 예측 모형을 구축하였다. 그러나 재무비율과 같은 회계 정보를 이용한 부도 예측 모형은 재무제표 결산 시점과 신용평가 시점 간 시차를 고려하지 않을 뿐만 아니라 해당 산업의 경제적 상황과 같은 외부 환경적인 요소를 반영하기 어렵다는 한계점이 존재하였다. 기업의 부도 여부를 예측하기 위해 정량 정보인 재무 변수만을 이용하는 것에 한계가 있음에도 불구하고 정성 정보를 부도 예측 모형에 반영한 연구는 아직 미흡한 실정이다.

본 연구에서는 재무 변수를 이용하는 기존 부도 예측 모형의 성과를 개선하기 위해 빅데이터 기반의 정성 정보를 추가적인 입력 변수로 활용하는 부도 예측 모형을 제안하였다. 제안 모형의 성과 향상은 정성 정보를 예측 모형에 통합시키기에 적합한 형태로 정보의 유형을 변환시킬 수 있는가에 따라 달려 있다. 이에 본 연구에서는 정성 정보 처리를 위한 방법으로 빅데이터 분석 기법 중 하나인 텍스트 마이닝(Text Mining)을 활용하였다. 해당 산업과 관련된 경제 뉴스 데이터로부터 경제 상황에 대한 감성 정보를 추출하기 위해 도메인 중심의 감성 어휘 사전을 구축하고, 구축된 어휘 사전을 기반으로 감성 분석(Sentiment Analysis)을 수행하였다. 형태소 분석 등을 포함한 텍스트 전처리 과정을 거쳐 감성 어휘를 추출하고, 각 어휘에 대한 극성 및 감성 점수를 부여하였다. 분석 결과, 전통적 부도 예측 모형에 경제 뉴스 데이터에서 도출한 정성 정보를 반영하는 것은 모형의 성과를 개선하는 것으로 나타났다. 특히, 경제 상황에 대한 부정적 감정이 기업의 부도 여부를 예측하는 데 더욱 효과적임을 알 수 있었다.

주제어 : 부도예측, 빅데이터 분석, 텍스트 마이닝, 감성 분석, 인공신경망

논문접수일 : 2016년 5월 3일 논문수정일 : 2016년 5월 31일 게재확정일 : 2016년 6월 13일
원고유형 : 일반논문 교신저자 : 신경식

* 이 논문은 2013년도 정부재원(교육부)으로 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2013S1A3A2054667).
** 이화여자대학교 경영대학, E-mail : namok.jo@gmail.com
*** 교신저자 : 신경식
이화여자대학교 경영대학
52 Ewhayeodae-gil, Seodaemun-Gu, Seoul, 120-750, Korea
Tel: +82-2-3277-2799, Fax: +82-2-3277-2776, E-mail: ksshin@ewha.ac.kr

저 자 소 개

조남옥
이화여자대학교에서 빅데이터 분석 기법을 경영분야에 적용하는 연구로 경영학 박사 학위를 취득하였고, 현재 이화여자대학교 경영대학 경영연구소 박사후과정 연구원으로 재직 중이다. 주요 연구분야는 지능형 의사결정지원시스템, 데이터 마이닝, 텍스트 마이닝, 빅데이터 분석 및 비즈니스 애널리틱스 등이다.

신경식
현재 이화여자대학교 경영대학 경영학부 교수로 재직 중이다. 연세대학교 경영학과를 졸업하고 미국 George Washington University에서 MBA, 한국과학기술원(KAIST)에서 인공지능, 지식기반 시스템 등 지능형 기법을 경영분야에 적용하는 연구로 경영공학 Ph.D.를 취득하였다. 주요 연구분야는 데이터 마이닝과 비즈니스 인텔리전스, 빅데이터 분석/비즈니스 애널리틱스, 인공지능 응용과 지식공학 등이다.