Diagnosis of Corporate Insolvency using Massive News Articles for Credit Management

*Note: Sub-titles are not captured in Xplore and should not be used

Hoon JIN
Big data analytics team
Xinapse co.,Ltd.
Seoul, South Korea
bioagent@xinapse.ai

Jeoung-Pyo Hong
Big data analytics team
Xinapse co.,Ltd.
Seoul, South Korea
jone.doe@xinapse.ai

Kang-Ho Lee
Big data analytics team
Xinapse co.,Ltd.
Seoul, South Korea
kangholee@xinapse.ai

Dong-Won Joo
Big data analytics team
Xinapse co.,Ltd.
Seoul, South Korea
joo@xinapse.ai

***Abstract*—In the aftermath of the 4th Industrial Revolution, AI and Big data technology have been used in various fields in South Korea, and the techniques are being applied to and complemented in various service fields which were implemented without them before. Especially, in order to secure credit stability for borrowed companies from financial institutions and to preemptively respond to the risks about—by means of online news articles and SNS data—the attempts to forecast the possibility of insolvency and adopt them into actual business are actively conducted by major domestic banks. In this study, we describe several analytical methods, outputs, and problems that are encountered during the processes of developing the unstructured text-based prediction system to detect the possibility of corporate insolvency—which ordered by a national government bank and discuss related issues with a real case. As a result, we have implemented an automatic tagger program for labeling largely unlabeled articles, and newly devised a prediction algorithm of the possibility of corporate insolvency. We achieved the accuracy of 92% (AUC 0.96) in aspect of performance and the hit ratio of 50% among the number of predicted 26 candidates that have the possibility of insolvency. Thus, the result of our study is revealed to be complementary to the financial data analysis sufficiently in performance, but yet have several limitations such as data coverage, reliability, and the characteristics of Korean language.**

Keywords—Unstructured data, corporate insolvency, text mining, machine learning, big data, news articles, labeling, prediction

I. Introduction

Interests in Artificial Intelligence and Big data related technologies are now increasing day by day in Korea, especially after the match of Go game between Google's AlphaGo and Lee Sedol in 2016. As a result, the wave of the fourth industrial revolution, since 2015y when the government had leading a boom by taking a global trend, and it has leveraged the development of those technologies that can be applied to a variety of industrial fields and pull a leading adoption. Among them, to secure the creditworthiness of banks from the companies who borrowed funds and to preemptively respond to the risks if maybe occur, several attempts has begun actively. One of the mainstreams of them is to develop the system forecasting the possibility of insolvency by using unstructured data such as online news articles and SNS, and adapt it to practical business works [6]. For an example, A local bank A has established the company diagnosis system by utilizing Big data and machine learning for risk management aiming to open in the first half of 2018. It provides the analyzed outputs on the screen after extracting bankruptcy events and keywords per each failure pattern from news articles. As another example, bank B has established a risk-based diagnosis system focusing on financial transactions and inter-company's risk transition with the aim of opening in May 2018 and monitors the occurrence situation of negative keywords on the news media to the borrowed companies by using unstructured text data. In the case of bank C, it established an early warning system to improve the credit monitoring system by using new techniques with the goal of opening in June 2018. It provides analyzed results using online news articles for credit management of the borrowers and lets it refer to company and industry news on loan tasks. In this paper, we describe various analytical methods, results, and related problems conducted in the process of forecasting the probability of corporate insolvency, which ordered by a national bank D, and discuss about its effectiveness in terms of performance and applicability.

The paper is organized as follows. Section 2 reviews background researches on bankruptcy prediction based on text mining methods. Section 3 describes our strategic analytical methodologies. Section 4 presents experimental results about data collection, refinements, preprocessing, evaluation and diagnosis of corporate insolvency. Section 5 describes conclusion from obtained results and discuss the special features of our research and limitations of text mining methodologies including ours.

II. Backgrounds

Many studies have actively been carried out since 2010 to try to predict the possibility of insolvency by using statistical or AI originated techniques for credit management of companies. In fact, such attempts with similar goals using statistical techniques have begun in the 1960s [5]. But, at this time, many studies for finding the possibility of insolvency formed the mainstream by utilizing only financial data, for an example, structured data recorded in numerical form. Especially in the late 20th century, early warning systems began to be developed around financial institutions [4]. And various statistical techniques and some AI algorithms have been used for the purpose of that. In the prediction of insolvency which conducted at before and after 2010, when the applicability of news articles

978-1-5386-7789-6/19/$31.00 ©2019 IEEE

was increasing—which had been issued previously, a lot of studies were reported. The demand for the utilization of unstructured data is caused by the following 3 points; (1) the increase of the number of SNS users; (2) the rapid increase of SNS data due to the spread of smart phones; (3) the increase of the number of news media led to the exponential increase of the amount of text data. [10] reports that text mining approaches are comparatively rare due to the difficulty of extracting relevant information from unstructured data. And it has goals to briefly describe the prototypes developed thus far. In the study of [12], the authors propose the method for sentiment analysis using news articles. 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. [3] refers that textual statements contain not only the effect (e.g., stocks down) but also the possible causes of the event. [9] reports that they used some general guidelines when manually classifying articles. For examples, mergers were generally considered positive because they indicate companies have cash on hand. Technology and general interest articles were considered neutral as they are not directly related to stocks. Lawsuits were generally considered negative, as was corruption. Rising interest rates were considered negative and declining interest rates were considered good because they indicate more cash in the general economy

III. Design and Methods

A. Overall Processes

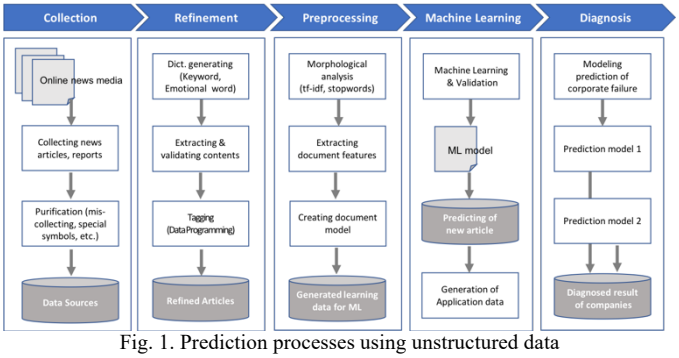

Fig. 1. Prediction processes using unstructured data

Fig. 1 shows the entire process that we designed. In general, the machine learning processes are known as <collection - preprocess - ML classification – evaluation - prediction>. However, unlike the experiment in laboratories, in our case the real data is collected and used, the collected data should be refined and labeled for building as learning data for machine learning analysis (or prediction). To this end, we inserted a new process called 'refinement' between the collection and the preprocessing. In addition, after the prediction of news articles, we added another new process which called 'diagnosis' in order to predict the insolvency of the company.

B. Collecting and Refining

Collecting process includes a 'refinement' step for cleansing as well as collecting news articles from about 120 online news media per each company. In this step, although the news data is collected by using the 'company name' and 'supplementary words' used for the detailed search, it consists of several steps for removing the articles that collected without relevancy of the company or having empty contents due to various reasons, or that were only written in a table or written as too difficult to perform an effective analysis because of many special symbols like controlled characters. As shown in <Table 3>, we collected news articles from predefined news media out of domestic internet sites.

In this study, we have tried to develop as considering systematic factors for practical use in financial institutions, unlike other studies mentioned above. It needs trying to derive the possibility of insolvency by using the newly predicted articles including the trained data. Therefore, we were forced to set the collection period to compare, analyze, combine, and supplement the analyzed result of the unstructured data with that of the structured. For the purpose of it, we designed to collect articles for four years including the year and month that the defaulting borrower went broken, and for the same period until December 2017 in case of the normal borrower. For example, we collect articles about company A from 2013 to 2016 if it went bankrupt in January 2016. If company B was bankrupted in August 2014, articles from 2011 to 2014 will be collected. As for the normal borrowers, because they have not been reported any events related to bankruptcy, we collect all the articles from 2014 to 2017 and use them for machine learning. And news articles from January to April of 2018 are collected additionally for prediction and selected as test data. To ensure stable experiments, the number of borrowers are decided to be about 800 at the same number of insolvent and normal borrowers.

C. Labeling for learning data

Unlike the data used for laboratory research, the actually collected articles from online news media are not equipped with labels for classification or prediction, and the classification labels can be changed according to the analyzed purposes. So, through the analysis of various economic issues and emotional information of the company appeared in news articles, we determine whether each article is positive or negative and set the label to predict the possibility of individual borrower's failure. In order to carry out labeling, relevant domain experts should read and check the articles one by one, and then judge the article to be positive or negative through logical evaluation. But labeling is usually done manually by human experts (or the users), which is a labor intensive and time-consuming process [2]. Therefore, after reading about 1 million articles in a short period of time, it is not possible to judge clearly but it is difficult to employ sufficient number of experts. Therefore, as a realistic alternative in the previous research, we employed a small number of intern staffs, fully explained the process and let them perform the task [5]. But because individual news articles have ambiguous expressions or technical terms in the content, it was actually impossible for a non-expert to judge them into positive or negative, after clearly understanding jargon and context. In addition, there were few cases in which the evaluators had given different results, and there were many cases in which even the same evaluator gave different results depending on when they were reading.

Such ambiguities and inaccuracies in the labeled results have a significant impact on the performance of prediction. No matter how good the algorithms are, or the latest techniques applied, vagueness in labeling of the training data cause to generate false training models, which in turn may lead to shaky and unreliable situations. Therefore, we have developed a rule-based data program to solve this problem. Fig. 2 shows the operation mechanism of the module automatically that classifies them as positive or negative, together with evidences of the judgement using the predefined sentimental dictionary for labeling [1].

The data dictionary is shown in Table 1. It consists of positive or negative keyword sets, positive or negative emotional word sets, and other important word set that will be explained in the section 3-D. In order to construct keyword and emotional word sets, a relevant domain expert needs to read a large amount of news articles belonging to the training data, understand the terms displayed in the individual article, then should evaluate and add the label with it. Generally speaking, the keyword set consist of noun-typed terms, and the terms are used importantly in evaluating the positive or negative property of an article in relation to the economy or the corporate credit management. The emotional word set have characteristics of verbal or adjective meaning and form, but are not limited to specific domains. The handcrafted dictionary like the above is used to perform labeling on news articles collected for all borrowers. In most cases, evaluation process does not meet the inevitably needed consistency for generation of the predictive model because news articles contain a variety of news (events, stocks, and etc.) about multiple companies rather than just a single news story about a single company. Also, reporters hardly presume the availability of individual news article as data when writing it. Thus, we designed a method of only extracting the content referring for the target company on individual article, and extracted only the company name based the partial contents, which are paragraphs or sentences, through repeated operations using the dictionary (Fig. 2).

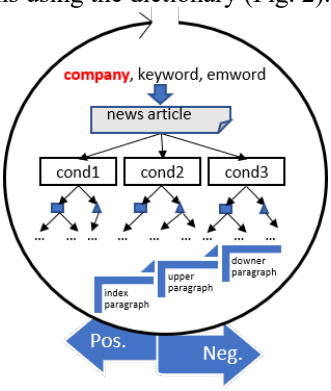

Fig. 2. A diagram of tagger's operation for each news article

However, since a person (or an expert) does not directly read and evaluate the articles, it may cause distrustfulness to the results in reliability as well as accuracy. Therefore, the validation process needs to be performed on the document sets which were evaluated to be positive and negative by the similarity verification test per individual article. For example, when there exist 100 documents—which were evaluated to be positive or negative, respectively—if we measure the similarity for a single document with the rest of 99 repeatedly, the degree of the average similarities from 0 to 1 can be calculated. Because they are just derived from mechanically operated similarity scores, it may not be true that the document with score 1 means that it belongs to the right category, which is positive or negative with the accuracy of 100%, and the document with zero similarity is irrelevant with the opposite category. Therefore, it is required to perform the additional validation process on documents which are found to be similar with.

TABLE 1. The number of terms which used in the document model of the data dictionary

Dictionary Type		#Terms	Example
Keywords	Positive	84	restriction of price range, a sharply upward trend …
	Negative	152	inspection, financial difficulty, imprison, …
Emotional Words	Positive	135	bullish, overcome, high …
	Negative	75	decrease, founder, encounter, …
Important Words	TF-IDF	254	Parliament, Shipbuilding industry, …

D. Data Preparation and Document Model Generation

In the preprocessing process, the entire document set should be divided to the training (learning + verification) and the prediction (test) set for machine learning. After several Korean morphological analyses were performed, a feature selection process is followed. Then the document model would be generated for machine learning operation. For morphological analysis, we used KoNLPy, which is now being provided as an open source packages for Python-based Korean morpheme analysis [14]. KoNLPy has multiple libraries such as Kkma, Komoran, Hannanum, Twitter, and Mecab. We conducted a preliminary test on Hannanum, Twitter, and Mecab, and found that Mecab was the most suitable for our purpose and expecting performance. We also registered about 30,000 companies and human names in Mecab dictionary for enhancing the purity of document features, which to be filtered. And including them, total number of 40,000 stop words were built and used. Selection of document features greatly affects the performance of prediction results. Nevertheless, the glossaries constructed as the above are used to be important for human judgment, but they may not be important in mechanical analysis like machine learning. So, it is necessary to add terms that can be used mechanically to improve the performance of machine learning. The adding word list, which are named 'important words', consists of the terms that were not appeared in the keyword or the emotional word sets among the glossaries, which are highly ranked through TF-IDF operation by using the documents belonging to the final set of training data after the preprocessing. Finally, we designed a set of about 2,000 document features using the data dictionary composed of five-words sets.

E. Machine learning and Prediction

The hot wind of the 4th industry revolution, by applying artificial intelligence technique, not only derives decision-making results which are difficult for human being to handle but also increased the number of studies that try to analyze and bring diversity in application areas. In the early stage of the project, due to the influence of Google AlphaGo, we are forced

to analyze and predict new articles and forecast the possibility of corporate failure by accepting deep learning technique. However, it is a well-known fact that deep-learning technique is not enough to provide sufficient evident for interpretation of the results and has effects only when the diversity and quantity of data are sufficiently ensured. As a result, machine learning can be more effective than the former and is preferred by our team. There are various kinds of algorithms in machine learning such as using numerical, probabilistic or evident-based model. Among them, Naïve Bayesian is traditionally known to perform well in classifying news data [11]. In recent years, support vector machine, which is applicable to both continuous and discrete data, and also known to be superior in accuracy, is often used. In addition, there are Decision Tree suitable for providing grounds of interpretation and Neural Networks having opposite characteristics to Decision Tree but known to provide relatively higher accuracy. Most widely known machine learning frameworks include TensorFlow, Weka, Caffe, Torch, Theano, Microsoft CNTK, Keras, Scikit-learn and extension modules provided by several mathematical or statistical packages such as Matlab or R [13]. TensorFlow has recently become the most popular for deep learning and Weka is frequently used for academic and research purposes since 2000y. As considering the overall situation, we have decided to use a python-based framework to build AI-based service system as well as to prepare future scalability and develop practically.

The trained result of positive and negative news articles that were collected and refined for the bankrupted and the normal borrowers in the previous step is generated as the learned model. At this time, it should be managed with file unit, considering that news articles that are continuously collected over time will be regularly predicted and classified and reflected to the model itself. The results of the predicted new articles and the trained ones should be used to derive the application data which necessary for generation of the prediction model in the next process. The application data needs to be made for including the article and the word frequencies to the company, and at the same time, considering together with the amount of time change.

F. Diagnosis of Corporate Failure

By using the ratios of the negative articles and the keyword aggregation scores generated through the trained and the predicted data, it can be derived from the probability of insolvency of the company. In case of the bankrupted borrowers, unlike the normal, the numerical values are recorded monthly per company and normalized in consideration of the frequency of the negative articles and the occurrence frequency of the terms in the handcrafted data dictionary according to the bankrupted time point, and then the dependent variable, in other words, the class value should be set to 1(bankruptcy) or 0 (normal). By preliminary experiments, the term aggregation scores per both company and month can be calculated. Then, the terms generated by each company and month are calculated as the '+' sign in case of positive and '-' sign in case of negative, considering the frequency, and finally the monthly representative value is expressed as the sum value.

In terms of algorithms, the ratio of negative articles is influenced by the progress of insolvency, and the frequency of occurrence increases as the month in which the bankruptcy occurs closer in most cases. However, in case of the term aggregation value, as the bankruptcy progresses, the increase or decrease of it varies not depending on whether the amount of the article but whether or not exists in the dictionary made beforehand. Therefore, we designed the continuous typed algorithm to apply in the case of using the ratio of negative articles to the prediction model 1 and select the model 2 in case of using the term aggregation ratio and apply the discrete typed algorithm.

TABLE 2. Definition of prediction models for the possibility of insolvency

- Model 1 : Explanatory variables – values of the frequency ratio of negative articles per month,
- Model 2 : Explanatory variables – values of the summed ratio of all the positive and negative terms per month,

(Both of the dependent variable - bankruptcy/normal)

IV. Experiments and Results

As shown in the below Table. 3, the number of collected news data is about 500,000 bankrupted borrowers and 600,000 normal borrowers. However, through the refinement process, each of 66.2% and 38.5% articles were removed compared to the collection and only 29.3% and 54.1% articles were used for training respectively.

TABLE 3. The state of data collection

Process	Bankrupted borrowers	Normal borrowers
# of companies	About numbers of 800	About numbers of 800
Collection	449,786	622,901
Refinement	151,983	382,953
Labeling (with validation)	132,053	337,388

The training data means that it has passed through both labeling and validation. The trained data is composed of the articles belonging to the bankrupted and the normal, and each of it includes the positive or the negative content. Fig. 3 shows the distribution of the number of the classified articles that have passed through the refinement process. According to the composition of the training data, the number of articles of the normal borrowers was more than twice as many as the number of the bankrupted. In addition, the number of positive articles in the bankrupted borrowers is about twice that of the number of negative articles, and about three times in the normal borrowers. The application data shows that the amount of articles increases as the time comes closer to the recent, and that even in the case of bankrupted borrowers, the number of positive articles is much higher than in the opposite case. 'Neutral' refers to the case where the judgment of positive or negative is not clear. And in the case of 'non-judgment', if the corresponding company name does not exist properly in the article, or no keywords or emotional words have appeared to evaluate to be positive or negative.

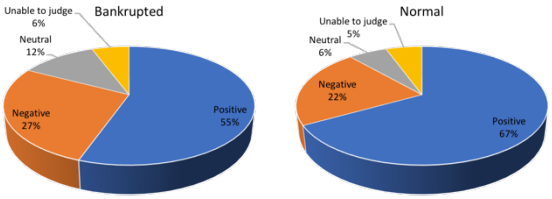

Fig. 3. The distribution ratio of labels of the articles in the bankrupted(left) borrowers and the normal (right)

In order to generate the document model using the confirmed training data, we chose 2,000 terms as described above, and then attempted several experiments by using man-made feature reduction (see the 1st~2th of Table. 4) and by using feature reduction through heuristics (see the 3st~4th of Table. 4). However, due to the problem that the performance was lower than the expectation value, the heuristic and recursive feature elimination algorithm (RFE) were used together instead of using single method. As switching the number of features from 100 to 1000 per 100 units, we found that the support vector machine showed the best performance at about 700 features.

TABLE 4. The performance of machine learning prediction in experiments

Exp.	Learning data (Ratio of train-val.)	Algorithm/ # of features	Performance
1	1th collection (9:1)	Man-made, 500	SVM, ACC- 65%, F1 score-0.743
2	1th collection (9:1)	Man-made, 1000	SVM, ACC-70%, F1 score-0.76, AUC-0.77
3	1th collection (9:1)	Heuristic, 640	SVM, ACC-75%, F1 score-0.8, AUC-0.83
4	1th collection (9:1)	Heuristic, 644	SVM, ACC-77%, F1 score-0.81, AUC-0.85
5	2th collection (9:1)	RFE, 101	MLP, ACC-71.9%, F1 score-0.78, AUC-0.79
6	2th collection (9:1)	RFE, 500	C4.5, ACC-80.2%, F1 score-0.84, AUC-0.89
7	3th collection (2:1)	RFE, 700	SVM, ACC-92.1%, F1 score-0.947, AUC-0.96
8	3th collection (7:3)	RFE, 700	SVM, ACC-93.8%, F1 score-0.957, AUC-0.96

In the process of diagnosing the possibility of insolvency, we devised model 1 and model 2 using the application data to predict the possibility of insolvency, as introduced in the previous section 3. Considering characteristics of the data, model 1 uses SVM and Gradient Boosting, and model 2 uses Decision Tree and Random Forest. In order to ensure the stability of the results in the process of developing the prediction models, we selected only companies having at least 100 articles for four years, and then, about 160 companies, which were only 10% of them, were identified as the experimental subjects. Due to small amounts of training data, we classified it by company size, listed company and bankrupted / normal borrower and experimented by applying the leave-one-out cross-validation (LOOCV) algorithm for learning and prediction. Fig. 4 is a visualization of the probability of insolvency for a specific normal company. The dashboard system was implemented similar to it. The blue line represents the model 1 using the ratio of negative articles, and the red line represents the result of model 2, which uses the terms aggregation ratios. Since the probability values of model 2 are discrete, all the probability values predicted by the four algorithms are calculated into the average only for the cases exceeding 0.5, and then confirmed as the diagnosis result of the probability about corporate insolvency.

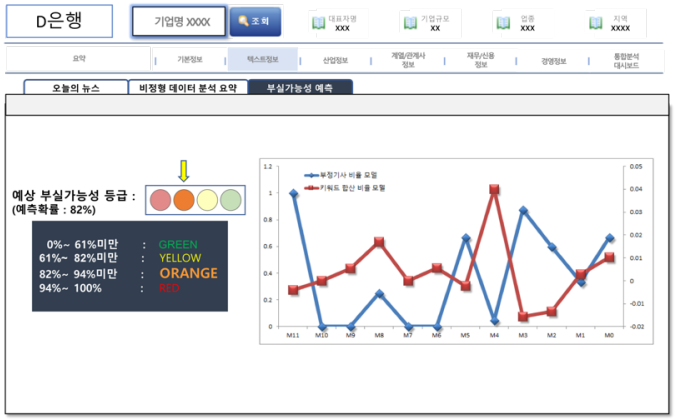

Fig. 4. Dashboard: The diagnosis result of the probability about corporate insolvency for company D in a specific month

In fact, it is unrealistic to use the previous four years of articles for the target borrower at the current time when we are going to apply the prediction function of the corporate insolvency to the credit management system. Therefore, we revised the model with only the previous 1-year amount, and the result was derived as shown in Table. 5. It shows some of the result that applied the window sliding method from May 2017 to April 2018 for normal borrowers who have not bankrupted in the past. Looking at Table. 5, instead of numeric values, test labels such as 'Green', 'Yellow', 'Orange' and 'Red' are recorded, that were the outputs of structured data analysis using financial information separated from our study. So referring to it, we attempted to classify the outputs of the prediction probability of the candidate company by 4 colors. It is assumed that we adjusted the numerical range to the characteristics of the unstructured data according to the ratio of the number of external corporations by applying the International Financial Reporting Standards of External Audit of Stock Companies. The output of the diagnosis of a specific company is calculated to be likely to insolvency in the present month when the probability of prediction is 0.61 or higher. If it is lower than 61%, the label is 'Green'. But it exceeds 61% and is lower than 82%, 'Yellow'. If it exceeds 82% and is lower than 94%, 'Orange', and the above of 94% is evaluated as 'Red'.

Finally, total 26 companies were predicted to be insolvent as of December 2017 and only 13 companies of them were found to be matched when compared with the outputs of the structured data analysis. Although it seems that only about 1.5 percent of the total number of 800 normal borrowers are matched, in reality, only about 600 companies have more than one article at least, and if we set a limit them to four years, there remains at most only 112 companies having more than 100 articles. Considering that bankruptcies are occurred mostly in middle and small sized companies, the number of those sized companies is 98. Therefore, the predicted 26 companies should be compared with 98 companies in practice, which represents about 28% of the total. Consequently, the fact the predicted 13 companies were matched means that it has the accuracy of 50%, so we can say that the analysis of the unstructured data have proved to be valuable sufficiently in stability and reliability. Also the result shows our strategy and methodology in this paper are effective.

TABLE 5. Diagnosis result of the probability for corporate insolvency per company

present time

테스트차주명	is_bk	201705	201706	201707	201708	201709	201710	201711	201712	201801	201802	201803	201804
	N	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	YELLOW	GREEN	GREEN	GREEN	YELLOW
	N	YELLOW	YELLOW	GREEN	YELLOW	GREEN	YELLOW	ORANGE	GREEN	GREEN	GREEN	GREEN	YELLOW
	N	GREEN	GREEN	GREEN	GREEN	GREEN	YELLOW	GREEN	YELLOW	GREEN	GREEN	GREEN	RED
	N	GREEN	GREEN	YELLOW	GREEN	ORANGE	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	ORANGE
	N	YELLOW	GREEN	GREEN	YELLOW	GREEN	GREEN	GREEN	ORANGE	YELLOW	YELLOW	YELLOW	YELLOW
	N	GREEN	GREEN	GREEN	GREEN	GREEN	RED	YELLOW	ORANGE	YELLOW	ORANGE	YELLOW	YELLOW
	N	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	YELLOW	GREEN	GREEN	YELLOW	GREEN	YELLOW
	N	GREEN	YELLOW	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	YELLOW
	N	YELLOW	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	ORANGE
	N	YELLOW	GREEN	YELLOW	GREEN	YELLOW	YELLOW	GREEN	ORANGE	YELLOW	YELLOW	ORANGE	YELLOW
	N	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	YELLOW
	N	ORANGE	RED	ORANGE	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN
	N	GREEN	GREEN	GREEN	GREEN	YELLOW	GREEN	GREEN	GREEN	GREEN	ORANGE	YELLOW	YELLOW
	N	YELLOW	YELLOW	ORANGE	ORANGE	YELLOW	YELLOW	YELLOW	ORANGE	YELLOW	ORANGE	YELLOW	YELLOW
	N	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	YELLOW	GREEN	GREEN	GREEN	GREEN	YELLOW
	N	GREEN	YELLOW	GREEN	YELLOW	GREEN	YELLOW	GREEN	GREEN	GREEN	YELLOW	YELLOW	GREEN
	N	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	YELLOW	YELLOW
	N	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	YELLOW	YELLOW	GREEN	GREEN	ORANGE
	N	GREEN	ORANGE	YELLOW	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	YELLOW	YELLOW	YELLOW
	N	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	ORANGE	RED	GREEN	YELLOW	GREEN	ORANGE
	N	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	YELLOW	YELLOW	YELLOW	GREEN	YELLOW
	N	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	ORANGE	GREEN	ORANGE	GREEN	RED
	N	GREEN	GREEN	YELLOW	GREEN	YELLOW	GREEN	GREEN	GREEN	YELLOW	GREEN	YELLOW	ORANGE
	N	GREEN	YELLOW	GREEN	GREEN	GREEN	ORANGE	YELLOW	YELLOW	YELLOW	GREEN	YELLOW	YELLOW
	N	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	ORANGE	GREEN	GREEN	YELLOW	GREEN	GREEN
	N	GREEN	GREEN	GREEN	GREEN	YELLOW	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	YELLOW
	N	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	YELLOW	YELLOW
	N	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	GREEN	YELLOW	GREEN	GREEN	ORANGE

V. Conclusion and Discussions

Using the unstructured data on online, attempts to analyze the economic and emotional information of the articles about companies in business, attempts to make prediction of stock price, marketing and credit management are spreading. However, a few cases have been known so far that such a system was developed and implemented for the implementation purpose.

In this paper, we reported a case study of the system implementation that predicts the probability of company insolvency by using unstructured big data in a Korean national bank, following the previous research [6]. We have found that it is difficult to apply many latest advanced algorithms or techniques, in spite of various technological advances in the fields of text mining—in order to develop the system that actually works in conjunction with the legacy system. Particularly, even trivial information beyond the experimental level of research or prototyping is not allowed to probabilistically approach with only vague possibilities in the financial sector, in which may cause economic loss and burden. Therefore, this study includes various detailed plans, attempts, and results to overcome several restrictions and application leveled limitations to develop realistic systems. The development of a tagger for labeling large volumes of unlabeled articles, and the design and development of models for predicting the possibility of corporate insolvency from the news articles are rarely discussed in previous studies. In views of performance, the result provides much higher accuracy and reliability than those other studies on the frontline.

Nonetheless, the system for predicting the possibility of insolvency using text-based unstructured data seems to be difficult to apply immediately or directly to the practice due to several issues and limitations. The first limitation is the problem of data coverage. Due to the nature of news articles or online data, large scaled companies or the companies with higher interests produce a large amount of related content, while in the opposite case, imbalances exist in the usability of the entire data due to the small data. It may even be difficult to derive analysis results due to insufficient training data. The second is the reliability problem of unstructured data analysis results. As the characteristics of the financial sector, the results of producing, processing, and analysis information related to creditworthiness of individual company should be accurate, predictable, and logically explainable as like numerical data calculation. However, unstructured text data are not easy to quantify, and even if it produces numerical prediction result, through the application of algorithms in the fields of text mining, ~~it~~ cannot be easy to apply to actual tasks, but also to understand the associated staff. The third problem is related to the specificity of Korean language, known as Hangul. Hangul belongs in ~~a~~ non-English cultural area and it has its own language system that operates through a combination of consonants and vowels that are different from other Asian countries. Thus, even if the same word occurs, the change in ending of a word or ambiguity occurs, and even if it is the same expression, it may have a completely different meaning from the original meaning of the word depending on the position of the tone and word class. It is also cultural characteristics of Korean language that the expression of the subject and the predicate are often used ambiguously and that there are many compound sentences frequently appeared in the sentences.

References

[1] A. Esuli and F. Sebastiani, “SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining”, In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC’06), pp. 417–422, 2006.

[2] B. Liu, X. Li, W. Lee, and P. Yu. “Text classification by labeling words”, In AAAI, 2004.

[3] B. Wiithrich, D. Permunetilleke, S. Leung, V. Cho, J. Zhang and W. Lam, “Daily Prediction of Major Stock Indices from textual WWW Data”, KDD-98 Proceedings, 1998.

[4] D. Olson, D. Delen and Y. Meng, “Comparative analysis of data mining methods for bankruptcy prediction”, ELSEVIER, Decision Support Systems, Vol. 52, pp.464-473, 2012.

[5] E. Altman, “Financial ratios, discriminant analysis and the prediction of corporate bankruptcy”, Journal of Finance, vol. 23, no. 4, pp. 589–609, 1968.

[6] H. Chen, P. De, Y. Hu and B. Hwang, “WISDOM OF CROWDS: THE VALUE OF STOCK OPINIONS TRANSMITTED THROUGH SOCIAL MEDIA1”, Review of Financial Studies, Vol. 27, pp. 1367–403, 2014.

[7] H. JIN, J. Hong and D. Joo, “The Prediction of Company’s Warning Sign through Unstructured data Analysis”, ICGHIT 2018 Conf., Thailand, Feb. 2018.

[8] H, Kim and S. Sohn, “Support vector machines for default prediction of SMEs based on technology credit”, European Journal of Operational Research vol. 201, issues 3, pp. 838–846, 2010.

[9] J. Zhai, C. Nicholas and A. Anand, “CS224N Final Project: Sentiment Analysis of News Articles for Financial Signal Prediction”, pp. 1-8, 2011.

[10] M. Mittermayer and G. Knolmayer, “Text mining system for market response to news: A survey”, Working Paper No 184, 2006.

[11] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. Trippe, J. Gutierrez and K. Kochut, “A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques”, ArXiv e-prints, arXiv:1707.02919, 2017.

[12] N. Jo and K. Shin, “Bankruptcy Prediction Modeling Using Qualitative Information Based on Big Data Analytics”, J. of Intelligence and Information Systems, vol. 22, issue 2, pp. 33-56, 2016.

[13] https://dzone.com/articles/11-open-source-frameworks-for-ai-and-machine-learn, Sep. 2018. Accessed.

[14] http://konlpy.org/ko/latest/, Sep. 2018. Accessed.