Journal of Theoretical and Applied Information Technology
31st August 2022. Vol.100. No 16
© 2022 Little Lion Scientific

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

KOREA GRAMMER-BASED SENTIMENT DICTIONARY DESIGN TO IMPROVE THE RELIABILITY OF OPINION MINING TECHNIQUES

JI HOON SEO[1]

[1]Institute of General Education, Incheon National University, Incheon 22012, Republic of Korea

E-mail: [1]jihoon@inu.ac.kr

ABSTRACT

The contemporary society is shifting to a paradigm that creates valuable data by utilizing various types of data. The value of such data is highly appreciated, and the vast asset serves as useful analysis materials in information warfare. Entering the 21st century, the amount of data is increasing geometrically. The more data collected from such big data, the higher accuracy of its analysis, and the more efficient data extracted. Big data analysis consists of structured and unstructured data analysis. One of the analysis methods of text-type unstructured data is opinion mining. The analytical method is advanced text mining, which determines the status of the given sentence as either positive or negative based on sentiment words to analyze reputation. This study draws out an integrated Korean-based sentiment dictionary algorithm to enhance the accuracy of reputation analysis when using Korean as sentiment words in opinion mining analysis. Since Korean grammar shows different characteristics in vocabulary formation compared to English, SWN used in English grammar analysis has limitations in its application to Korean. Thus this study aims to build up a Korean-style sentiment dictionary, thereby proposing a methodology to increase the precision of opinion mining analysis.

Keywords: *Big data, Opinion mining, Sentiment analysis, Sentiment data dictionary*

1. INTRODUCTION

As the concept of big data is firmly established following data increase, the big data analytical method is making up a great portion of the fourth industrial revolution. Since the revolutionary era analyzes and utilizes data collected from IoT and others to generate, process, and manage meaningful information, the data is considered as the future society asset. Furthermore, as IT computing technology develops and the distribution rate of the press, communication media, smartphone, and internet media experiences a rapid increase, various big-data mining studies continue to being unfolded [1][2]. Currently, large volume data is on rapid growth at home and abroad. Further, an explosive increase of data is expected due to Cloud Computing technology, Social Networking Service, Mobile Service, and IoT(Internet of Things) in the future. Such data is categorized into two large groups structured and unstructured, and the value and reliability of the analysis data depend on the availability of removing unnecessary data and drawing out meaningful data [3][4].

The 21st century faces increased analysis of unstructured data in a form of video, text, and image. Moreover, among them, opinion mining analytical method that analyzes sensitivity of text and judges its positive/negative status is becoming an issue. Opinion mining is a technology that utilizes user's opinions and sensitivity patterns to draw out three state positive, neutral, negative from an opinion on a certain object, and this requires the target object and an opinion regarding it. Also, the technique extracts words data representing positive or negative status from the text, recognizes an object and opinion on it, and measures the level of positiveness as the sum of patterns including opinions.

Its application is marketing or reputation analysis following consumer opinions. In case of opinion mining analysis for English words, you may use SentiWordNet (SWN), a half-supervised learning method established by Princeton University, to score sensitivity of the given words, and determine its status: negative/positive. SWN contains sentiment data that applied English grammar; significant sensitivity value of positivity/negativity was allotted to build up the dictionary for opinion mining analysis [5][6]. So far, it has been updated to version 3.1 and utilized in more than 300 research groups. SWN is characterized by synonym sets based on 147,278 English words including noun, verb, adjective, and adverb, and normalized so that positive word, negative word, and neutral word values make up one in total. As such, SWN is specialized in performing opinion mining analysis of English grammar [7].

Due to the structure of grammar rules, however, English and Korean grammar show differences in semantic, spatial, and noun-application, temporal, nuance, and syntax principles. Thus, SWN has difficulties in applications for Korean grammar. In this regard, the study intends to set up a Korean-style sentiment dictionary for a more efficient opinion mining analysis in Korean grammar and suggest a process and methodology to improve the accuracy of analysis.

2. RELATED WORKS

2.1 Sensitivity Analysis

Text of unstructured data consists of a topic of the context and opinion regarding it. We call the process of finding out the text topic "Topic Modelling"; it is the technology that provides automatic calculation of the topic from the given text Also, we refer to identifying the author's opinion on the topic as "Sensitivity Analysis" or "Sentiment Classification." With the technology, you may judge if a subjective or objective opinion is included in the text; in the case of the former, you may also find out if it is positive or negative [8][9].

2.2 Synonym Group Classification Technique

Synonym set classifies English words into synonym groups, provides general and concise definitions, and defines various semantic relations between vocabularies. The reason for applying this process is to combine the dictionary and thesaurus (synonyms and antonyms) to facilitate more intuitive usage, automated text analysis, and artificial intelligence applications. For instance, WordNet adopts synonym, and its version 3.1 established about 110,000 synonym sets such as nouns, adjectives, verbs, and adverbs [10].

2.3 Pointwise Mutual Information

PMI (Pointwise Mutual Information) is a polarity classification method that judges if the given two words are positive or negative [11]. Since it is easier to use or calculate and a more precise result comes out than others, the opinion mining industry widely uses it to classify given words. Its equation shows how closely word1 and word2 are connected based on probability theory.

$$\text{PMI(word1, word2)} = \log_2 \frac{p(word1,wor\quad)}{p(wor\quad)\,p(wor\quad)} \quad (1)$$

Web-PMI (Web-Pointwise Mutual Information) refers to partially modifying the PMI calculation equation so that the user may judge polarity through the number of documents searched by googling a word on the internet, and its equation is as follows;

$$\log_2 \frac{\frac{1}{N}hits(wor\quad,word2)}{\frac{1}{N}hits(word1)\,\frac{1}{N}hits(word2)} \qquad (2)$$

Here, N represents the total number of the documents, while hits(word1) shows the number of those containing word1. In other words, the equation draws out the final value calculated like this. First, divide the number of those containing both word1 and word2 with the total number of documents, N. Next, divide the number of those containing word1 or word2, respectively, by the total number of documents and multiply the two. Finally, divide the former with the latter value to draw out the final one.

2.4 Text Mining Technique

Text mining is a technique that aims to extract and process valuable information based on processing technology of natural languages, which are unstructured data. When collecting unstructured data, the data storage inevitably keeps both meaningful information and unnecessary ones; in fact, 8 out of 10 are the latter. Thus, text mining technique extracts significant information among massive text corpus, identifies its relations with other information, and categorizes each text, drawing out more results than simple information retrieval [11]. With large-volume language resources and statistical and regular algorithms, computers use human languages and find hidden information; its applications range from document classification, document clustering, information extraction, to document summarization. The related fields are opinion mining, which is sentiment analysis.

2.5 Cluster Analysis Technique

Cluster analysis is used to find similarities between entities and group them at last. For example, Twitter has mixed data on different interest areas. In cluster analysis, user groups can be classified according to such interest areas and hobbies. Cluster analysis has two methodologies; one is hierarchical grouping that puts together consecutively and the other is nonhierarchic one that predicts clusters before actually making groups.

2.6 Opinion Mining Analysis Technique

Opinion mining is also called "sentiment analysis" and broadly refers to natural language processing, computer linguistics analysis, and text mining. The theory of opinion mining has developed progressively since the early 2000s and many studies have been carried out by analyzing reputations and reviews of customers on electronic commerce.

Kim, Jinok (2011) focused on text polarity analysis to determine whether the author's opinions and attitudes revealed in Korean text are positive, neutral, or negative and proposed an automated opinion classification technique to distinguish positive and negative ones by generating text patterns through eight stages and adjusting a score of each pattern considering the context.

3. BUILD UP SENTIMENT DICTIONARY

3.1 Designing System Platform of Sentiment Dictionary

Generally, this study comprises of fundamental six steps in principle to perform opinion mining analysis;

1) Data Acquisition
2) Data Classification
3) Data Pre-Processing
4) Sentiment Dictionary Construction
5) Word Tagging
6) Data Analysis.

The overall system structure for the Korean-style sentiment dictionary has three stages storage server for data collection area, storage, and processing; NLP learning model for natural language processing, morpheme analysis, and pre-processing; Master server to establish an advanced Korean grammar-based sentiment dictionary and implement OM (Opinion Mining) analysis.

3.1.1 Storage Server

It is a sub-system to save the collected raw data and assist its processing. It performs crawling of unstructured data on the internet to gather it randomly.

3.1.2 Natural Language Processing Module

The module plays a role to classify unstructured text data and extract positivity or negativity from sentiment words to establish and keep a sentiment dictionary. The natural language processing module requires for dissemination of morpheme in sentences collected to extract sentiment words and performs Tokenizer process to separate the vocabulary in sentences into the minimum string form.

3.1.3 Master Server

Master server performs tagging of the polarity of sentiment words drawn out from the former step and sets up a management system for opinion analysis data. This study aims not to establish big data but seeks accuracy in derive opinions, and first proposes sophistication of sentiment dictionary as its method.

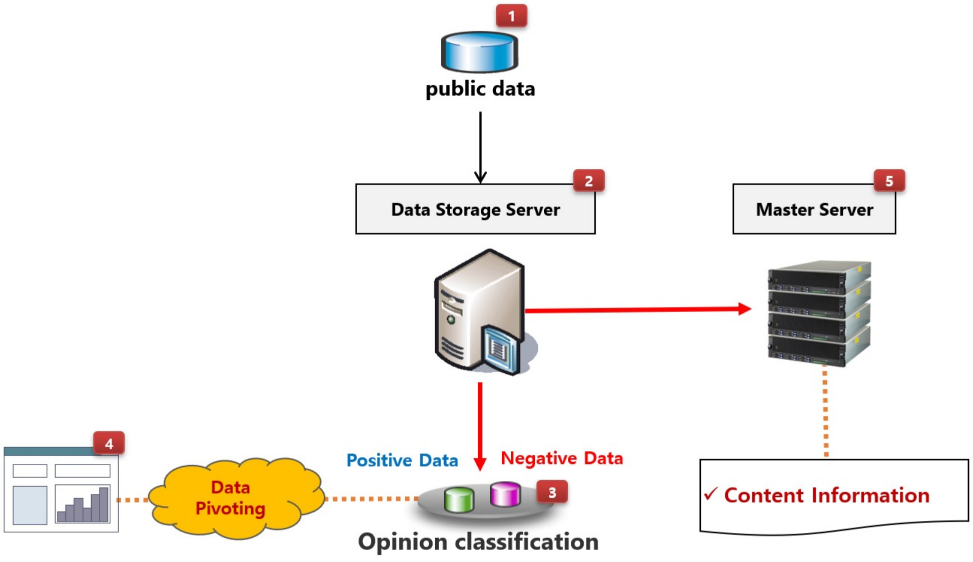

Figure 1: Overall System Configuration

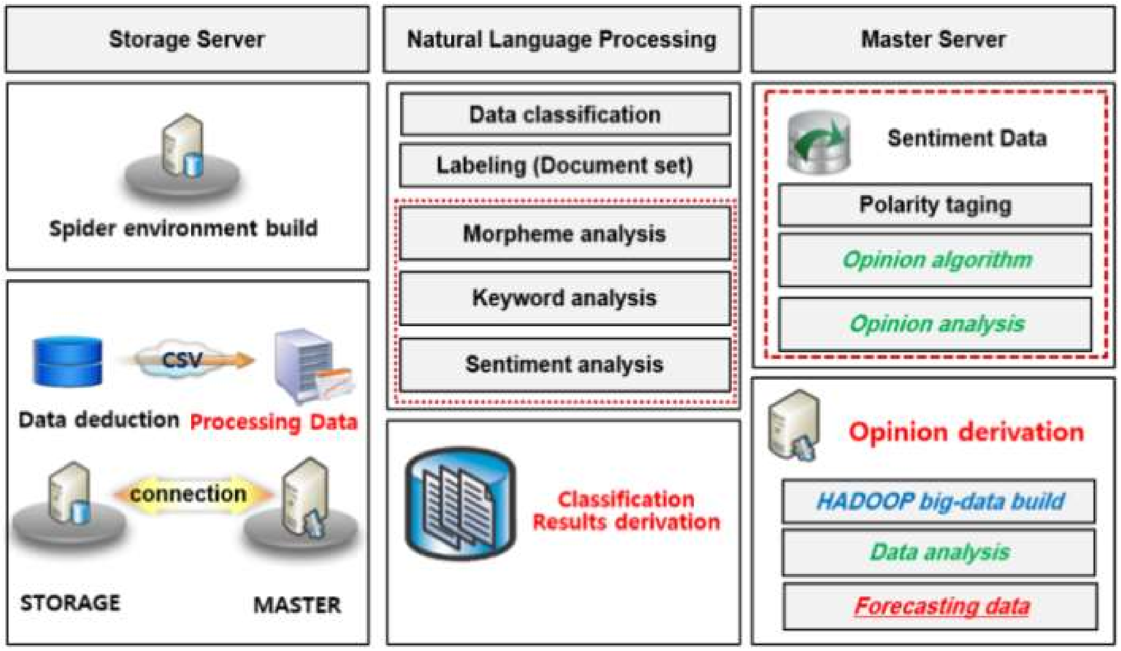

Figure 2: Sentiment Dictionary System.

3.2 Derivation of Characteristics Using NLP

Looking at examples of general appearance frequency of words in English, the highest group includes an article 'the' or conjunction 'and', and other words commonly used in most texts. Although such words naturally boast its high frequency in sentences, they cannot be interpreted as unique characters that connect context of the given text. To accurately find out words that clearly show characteristics of each text, you must extract unique words that occasionally appear, instead of generally derived words such as articles or conjunctions. Likewise, conjunctions are regularly used in Korean grammar, which hinders the accuracy of data and raises a question of reliability in considering characteristics of each sentence.

To eliminate such an unessential regularity of appearance frequency and to satisfy requirements, this study utilizes TF-IDF indicator. To calculate tf, it requires n_{ki} that represents frequency of word i appearing in text k, and N_k that shows the total number of words appearing in text k. Only then, you may draw out tf_{ki}, the tf value of word i in text k. idf_i, the idf value of word i is computed by using D, the total number of its appearance in the given text, and Di, the number of documents word i appears.

$$idf_i = \log\frac{D+1}{D_i} \rightarrow tfidf_{ki} = tf_{ki} \cdot idf_i \qquad (3)$$

Here, tf_{ki}, the tf value serves as an indicator that displays the frequency ratio of word i in a certain text k. Thus, when a word appears a lot in a given text, the tf value is bigger; on the contrary, when it turns up less often, the value becomes smaller. idf value displays the log value that divides $D + 1$ by D_I, which shows the number of text word i appears. Therefore, idf value is effective in excluding general words such as articles and conjunctions.

3.3 Filtering Pre-processing

Perform filtering to establish a sentiment dictionary based on data of potential sentiment word groups that were drawn out from characteristics of the collected data.

In Korean grammar, adopting pre-processing design for appropriate words and grammar rules must precede designing an opinion mining analysis model since one word has a different and abundant meaning in Korean grammar, which cannot be expressed in English grammar. Thus, when analyzing Korean grammar, we must apply new patterns and rules beforehand to enhance the reliability and accuracy of the analysis.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 1: Filtering Rules to Extract Potential Word.

* Applications for Word Filtering Rules to Build Up a Sentiment Dictionary
1. Remove special characters, English, and unnecessary vocabularies. 2. Get rid of meaningless words and one-letter-text. 3. Separate the identical word in a conjunction form to classify the essentials of sentiment words. 4. Classify homonym/heteronym. 5. In case of abbreviations and new-coined words, only use those registered on Wikipedia and Korean dictionary.

3.4 Antonym Correspondence Rule in Korean Sentences

In the case of Korean grammar, although there are many positive words in the entire sentence, there are cases where the composition of the entire vocabulary is reversed negatively due to antonyms. These elements differ depending on the purpose of the sentence or the context. In this study, a sentiment dictionary was constructed based on unstructured data related to stocks by designating one concept for the analysis target.

Table 2: Probability Rule for Derivation of Antonyms for Positivity word and Negativity word.

word	Positive Probability	Negative Probability
Rise	rose	after rising
	rise and	rose but
	while rising	rose for a while
	by rising	although rose
	rose and	rose but still
Fall	after falling	fell
	fell but	fall and
	fell for a while	while falling
	although fell	by falling
	fell but still	fell and

For example, the most frequently derived words in sentences related to stocks have the semantic central words “rise” and “fall.” The following [Table. 2] shows examples of words that are likely to be recognized as negative in the overall sentence in the vocabulary of “rise” and “fall.”

We may find abbreviations that have similar meaning with the above: “upper limit” and “lower limit.” In the case of the word “rise”, it includes various semantic elements such as "stocks rose", "stocks are expected to rise", "stocks rose but fell", etc. Also, there are antonyms that contain concluding words as negative sentences after positive words, such as “the stock rose but closed in a downtrend.” When analyzing these sentences in the overall context, although the word “rise” has a positive meaning and the “fall” has a negative meaning, there are elements for the antonym where the word “rise” shows a negative future-prediction in the context of the overall flow of the sentence.

In general, in the process of deriving a sentiment dictionary, there are cases where positive words are applied as they are, without considering the irony of the post-words in the sentence. In this study, words for positive vocabulary are considered as positive in the overall aspect, but a separate data dictionary is constructed by filtering predictive words to be converted into negative words.

3.5 Antonym Correspondence Rule in Korean Sentences

This study constructs a sentiment dictionary applicable to related themes. In general Korean sentimental classification, it is not easy to derive appropriate sentiment parameters for each field.

For example, while establishing a sentiment dictionary specialized for stock-related themes, a vocabulary generally perceived as negative can be recognized positively in the stock market. In the case of a “fall in interest rates”, it shows a negative tendency indicative of a recession, but a “fall in interest rates” in the stock market is generally regarded as a good news.

This is because stock prices and interest rates have a correlation with each other in the stock market. In other words, when interest rates and bonds show a downward trend, stock prices rise; conversely, when interest rates and bond prices rise, stock prices fall. This study also applied the same regularity to classify vocabularies by considering correlations and adjacencies such as the positive and negative effects of “falling interest rates” and “rising interest rates.”

Table 3: Examples of Relation for Positivity word and Negativity word

Word	Opinion	Relation	Division
fall	Negative word	Causal relation	stock fall
rise	Positive word	Causal relation	stock rise
falling interest rates	Positive word	Correlation	stock rise
rising interest rates	Negative word	Correlation	stock fall
Foreigners [Sale of Stock = Sell]	Negative word	Causal relation	stock fall
Foreigners [Purchase of Stock = Buy]	Positive word	Causal relation	stock rise

3.6 Resolving Antonyms Inverse Relation based on Connective Word

It constructs a secondary sentiment dictionary centered on connecting words that can identify the inverse relation of conjunctions from which irony is derived in Korean grammar, and contrast them.

The types of connective words in Korean grammar are largely divided into eight categories: direct relation, inverse relation, causal relation, correspondence relation, supplementary relation, affirmation relation, transition relation, and example relation. In this study, the form of the connective words with high lexical influence in the sentence is the inverse relation, and in this context, there are the grammar of the inverse form such as “but”, “however” and “though”.

Table 4: Irregular Predictive words

Keyword Word	Primary Irregularity	Secondary Irregularity
Rise (positive word)	did. (positive)	no change. (positive word)
	did, and (neutral) did but (neutral)	currently breathing space. (neutral word)
		fell (negative word)
	couldn't. (negative)	
Fall (negative word)	did. (neutral)	
	did and (neutral) did but (neutral)	no change. (negative word)
		currently breathing space. (neutral word)
		rose (negative word)
	didn’t. (positive)	

Due to the irregular grammar, the opinion tagging process in the sentence is merely a machine learning process, and in marking only positive and negative, the process only lowers the accuracy of determining whether the entire sentence is positive or negative. As a solution to this, this study regards the antonym as negative and applies the tagging rule that describes the whole word as neutral in the case of antonym grammar that occurs after the positive tagging word. For example, if an antonym exists before the positive tagging of “rise” in a word, the word “rise” is highly likely to suggest the opposite negative tagging by the antonym. Accordingly, the antonym before positive tagging is regarded as negative tagging, and "positive + negative" performs invalidation processing through neutrality declaration. On the contrary, the antonym that appears before the word for negative tagging nullifies "positive + negative" through positive tagging.

4. METHOD EXTRACTION OF OPINION DATA AND EVALUATION OF ACCURACY

In this study, opinion mining analysis was performed on 30 documents based on the online unstructured data extracted for one day. The method of calculating data on whether the entire document for one day is positive word(pos) or negative word(neg) based on data for one month is as follows;

$$OM\ pos\ and\ neg\ = \left(\frac{\sum_{i=1}^{n} 1day\ Document}{daily\ pos\ or\ neg\ data}\right) \quad (4)$$

The above is an average formula that calculates the positive or negative index of opinions generated in a day. 1day Document refers to the number of documents derived during one day, and daily Pos data refers to positive words generated from documents throughout the day. Based on the opinion index derived from the day, it is possible to identify the positives and negatives of the overall article about the day. As a sub-category, this divides opinion mining analysis for the document by deriving positive and negative data of unstructured data based on one document. The results of analyzing stock news data for one month through the construction of an integrated sentiment dictionary algorithm based on Korean grammar presented in this study are as follows;

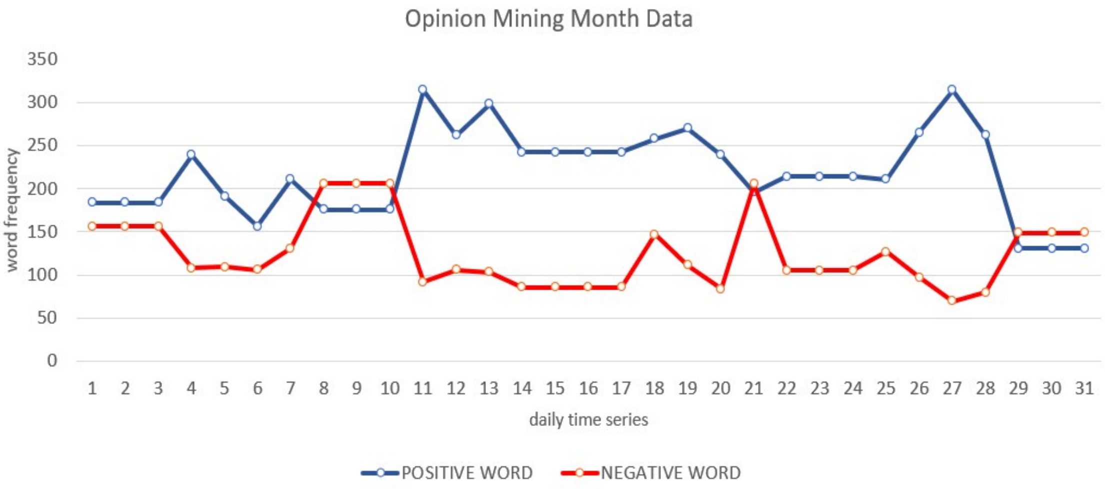

Figure 3: Opinion Mining Analysis on Month Data Before Algorithm Application.

Generally, it is the result of calculating 20 document data, and one data includes 30 news data. In the sentiment data before the rule application, the derivation for the positive accounted for 85 percent and the derivation for the negative accounted for 15 percent. The document data is the result of adding up each data on the opinion's positive and negative vocabulary derived from 30 news articles

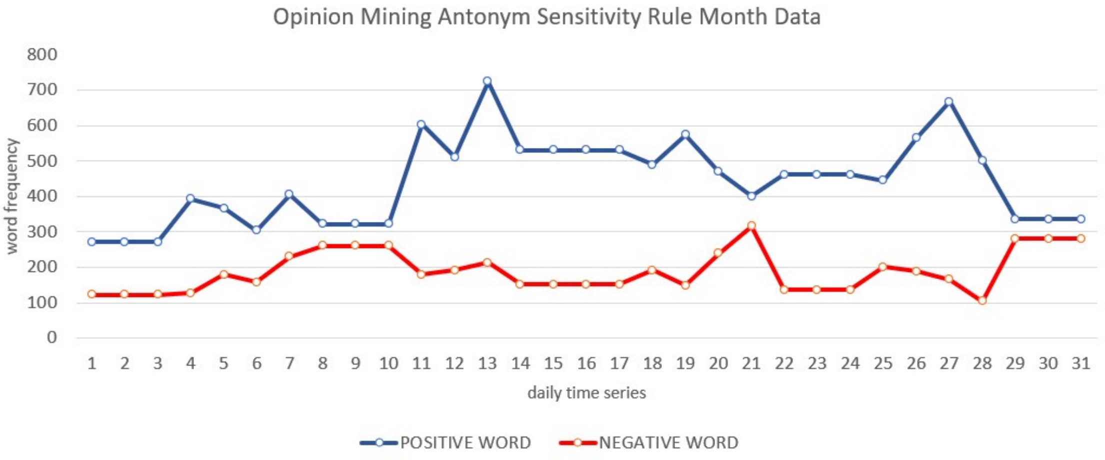

Figure 4: Opinion Mining Analysis on Month Data After Algorithm Application.

The sentiment data applied through this study are as follows. Although it did not show a significant difference from the value of opinion mining derived from the existing sentiment dictionary, the number of discrimination between positive and negative vocabulary increased more than twice, confirming that the accuracy was improved. occurred as a result of testing the deviation rate of positive and negative values for the algorithm applying antonym and inverse relation and the existing opinion derivation technique. It was found that the appearance rate of positive and negative vocabulary slightly increased.

Table 5: Performance Evaluation of Analysis.

Detailed Accuracy By Class	
Class	AVG.
TP Rate	0.974
FP Rate	0.029
Precision	0.975
Recall	0.974
F-Measure	0.974
ROC Area	0.982

5. CONCLUSION

In this study, in order to increase the accuracy of opinion mining for positives and negatives that fit Korean vocabulary, we designed a storage server through data collection and a master server for big data management, and proposed a new sentiment dictionary that applies association rules that fit the theme of a specific field. We designated stock data as a theme and collected a total of 20,000 online media news articles to build a sentiment dictionary. We built two sentiment dictionaries based on irony and inverse relation to vocabulary and joined them to build a new sentiment dictionary. In general, we performed an accuracy test to determine whether the whole sentence was positive or negative with respect to the number of news data published per day. We conducted an experiment based on the inference of how reliable the reputation of news data is through comparison of stock price data on a monthly basis and the accuracy of the methodology presented in this study. In the case of Korean, in order to understand the overall meaning of the vocabulary for natural language, improvement of ambiguity is required and an efficient methodology for syntax analysis is needed. In this study, the relation of connection words was interpreted as a method to improve the unilateral machine learning of the whole sentence according to the number of appearances of positive and negative words, and the result of improved accuracy was obtained through the appropriate sentiment dictionary. In the future, in-depth ontology analysis of semantic assignments, sentence patterns, and semantic inference constraints in Korean texts is needed. In addition, more sophisticated mining analysis should be performed based on Korean-based antonym form analysis and the appearance pattern of specific central words.

Acknowledgments. This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(No. 2020R1I1A1A01055323)

REFERENCES:

[1] Seo, JH, Choi, J.: Design for Opinion Dictionary of Emotion Applying Rules for Antonym of the Korean Grammar: JKIIT, 2015, Vol. 13, No. 2, pp. 109-117.

[2] Jo, HJ, Seo, JH,Choi, J.: OAR Algorithm Technology Based on Opinion Min- ing Utilizing Stock News Contents: JKIIT, 2015, Vol. 13, No. 3, pp. 111-119.

[3] Lee, YJ, Seo, JH,Choi, J.: Fashion Trend Marketing Prediction Analysis Based on Opinion Mining Applying SNS Text Contents: JKIIT, 2014, Vol. 12, No. 12, pp. 168-170.

[4] Seo, Ji-Hoon, Lee, Ho-Sun, Choi, Jin-Tak: Classification Technique for Filtering Sentiment Vocabularies for the Enhancement of Accuracy of Opinion Min- ing: International Journal of u- and e- Service, Science and Technology, 2015, Vol. 8, No. 10, pp. 11-20.

[5] Andrea Esuli, F. Sebastiani, SENTIWORDNET: A high-coverage lexical resource for opinion mining: Evaluation, 2007, pp. 1-26.

[6] E. Courses and T, Surveys. (2008). Using Sentiment SentiWordNet for multilingual sentiment analysis, IEEE 24th International Conference on Data Engineering Workshop, Cancun, Mexico, pp.507-512. 2008..

[7] Stefano Baccianella, Andrea Esuli, Fabrizio Sebastiani.: SentiWordNet 3.0: An En- hanced Lexical Resource for Sentiment Analysis and Opinion Mining: LREC, 2010, Vol. 10.

[8] Anindya Ghose, Panagiotis Ipeirotis, Arun Sundararajan, Opinion Mining Using Econometrics: A Case Study on Reputation System: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, pp. 416-423.

[9] Hsinchun Chen; David Zimbra.: AI and Opinion Mining: IEEE Intelligent Systems, pp. 74- 80.

[10] Ji-Hoon Seo and Kil-Hong Joo. (2019), Analysis of the Requtation of Domestic Software Education Using Big Data, The Society of Convergence Knowledge Transactions, Vol.7, No.4, pp.141- 148, 2019.

[11] Irfan Ajmal Khan, Junghyun Woo, Ji-Hoon Seo, Jin-Tak Choi, "Text Mining : Extraction of Interesting Association Rule with Frequent Itemsets Mining for Korean Language from Unstructured Data", International Journal of Multimedia and Ubiquitous Engineering SCOPUS , Vol.10 No.11, pp.11-20, 2015.