Thesis for the degree of
Doctor of Philosophy

The Business Impact of Social Media

- Sentiment Analysis Approach -

Supervisor: Prof. Yu, Song Jin

January 2017

The Graduate School
Korea Maritime and Ocean University

Department of Shipping Management

Kim, Dongwon

The Business Impact of Social Media

- Sentiment Analysis Approach -

by

Kim, Dongwon

A thesis submitted to the Department of Shipping Management in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in Business Administration at the Graduate School, Korea Maritime and Ocean University

Supervisor: Prof. Yu, Song Jin

January 2017

Approved by Thesis Committee:

Chairman:	Prof. An, Ki Myung	Ki myung An
Examiner:	Prof. Shin, Han-Won	Shinhan
Examiner:	Prof. Shin, Jae-Yeong	Jaeyoung Shin
Examiner:	Prof. Kim, Young Bu	
Examiner:	Prof. Yu, Song Jin	Song Jin Yu

Contents

3. Methodology

4. Experiment & Analysis

5. Conclusion

List of Tables

List of Figures

The Business Impact of Social Media
- Sentiment Analysis Approach -

Kim, Dongwon

Department of Shipping Management
Graduate School of Korea Maritime and Ocean University

국문초록

이 연구의 목적은 소셜 미디어에서 추출된 7개의 감성 도메인이 자동차 시장 점유율 예측에 대한 감성 분석 실험을 위한 데이터로서 적합한 지에 대한 신뢰성을 확인하고 고객들의 의견이 기업의 성과에 어떻게 영향을 미치는 지에 대하여 확인하기 위한 것이다. 본 연구는 총3단계에 걸쳐서 진행되었습니다. 첫 번째 단계는 감성사전 구축의 단계로서 2013년 1월 1일부터 2015년 12월 31일까지 미국 내 26개의 자동차 제조 회사의 고객의 소리 (VOC: Voice of the Customer) 총 45,447개를 자동차 커뮤니티로부터 크롤링 (crawling)하여 POS (Part-of-Speech) 즉 품사정보를 추출하는 태깅 (tagging)과정을 거쳐 부정적, 긍정적 감성의 빈도수를 측정하여 감성사전을 구축하였고, 이에 대한 극성을 측정하여 7개의 감성도메인을 만들었습니다. 두 번째 단계는 데이터에 대한 신뢰성 분석의 단계로서 자기상관관계분석 (Auto-correlation Analysis)과 주성분분석 (PCA: Principal Component Analysis)을 통해 데이터가 실험에 적합한지를 검증하였다. 세 번째 단계에서는 2개의 선형회귀분석 모델로 7개의 감성영역이 미국내 자동차 제조 회사 중 GM, 포드, FCA, 폭스바겐 등 총 4개의 자동차 생산 기업을 선정하여 이들 기업의 성과 즉, 자동차 시장점유율에 어떤 영향을 미치고 있는 지 실

험하였다. 그 결과, 우리는 4,815개의 부정적인 어휘들과 2,021개의 긍정적인 감성어휘들을 추출하여 감성사전을 구축하였으며, 구축된 감성사전을 바탕으로, 추출되고 분류된 부정적이고 긍정적인 어휘들을 자동차 산업에 관련된 어휘들과 조합하였고, 자기상관분석과 PCA (주성분 분석)를 통해 감성의 특성을 조사하였다. 실험 결과에 따르면, 자기상관분석에 의해서 감성 데이터에 어떤 일정한 패턴이 존재한다는 것이 발견되었고, 각각의 감성 영역의 감성이 자기상관성이 있으며, 감성의 시계열성 또한 관찰되었다. PCA에 의한 결과로서, 7개 감성영역이 부정성, 긍정성, 중립성을 주성분으로 연결되어 있음을 확인할 수 있었다. 자기상관분석과 PCA를 통한 VOC 감성 데이터에 대한 신뢰성을 바탕으로 2개의 선형회귀분석 모델을 구축하여 실험을 진행하였다. 첫 번째 모델은 주성분 분석에서 부정적 감성의 Sadness, Anger, Fear와 긍정적 감성도메인인 Delight, Satisfaction을 독립변수로 선정하고, 시장점유율을 종속변수로 선정하여 실행하였고 두 번째 모델은 첫 번째 모델에 주성분이 중립성으로 결과가 나온 Shame, Frustration을 독립변수에 추가하여 중립성을 띠고 있는 감성이 시장 점유율에 유의미한 영향을 미치고 있는 지를 확인하였다. 분석 결과, 각 기업 마다 시장점유율에 유의미한 영향을 미치는 감성들이 존재하고 모델 1과, 모델 2에서의 감성 영향력이 차이가 있음을 발견하였다. 본 연구를 통해, 데이터 상에 나타난 정보를 가진 감성이 과거 값에 기초하여 자동차 시장에서 변화를 수반할 수 있다는 것을 나타내고 있음을 확인하였다. 또한, 우리가 시장 데이터의 가용성을 적용하려고 할 때, 자동차 시장 관련 정보나 감성의 자기상관성을 잘 활용할 수 있다면, 감정 분석에 대한 연구에 큰 기여를 할 수 있을 뿐만 아니라, 실제 시장에서의 비지니스 성과에도 다양한 방법으로 기여할 수 있을 것으로 기대된다.

KEY WORDS: Auto-correlation; 자기상관관계, PCA; 주성분분석, Linear Regression Analysis; 선형회귀분석, Sentiment Analysis; 감성분석, Reliability; 신뢰성.

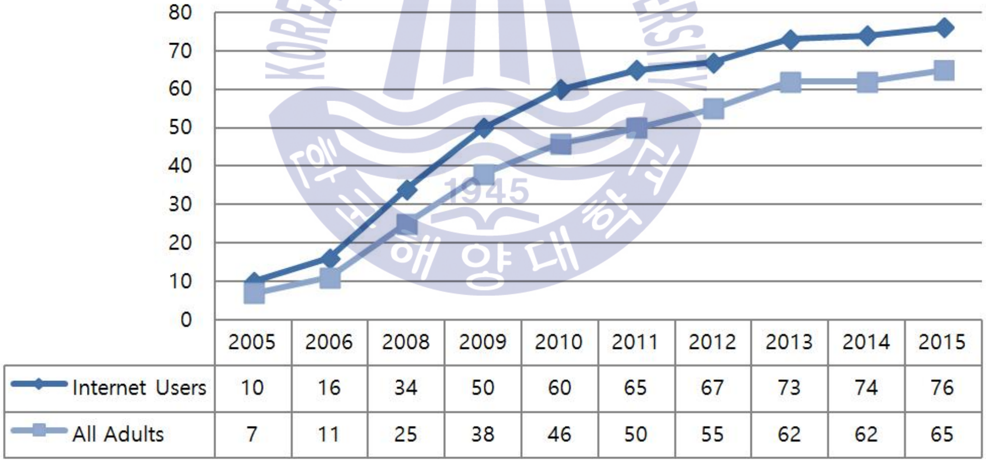
1. Introduction

1.1 Back Ground

In today's challenging business world, companies want to differentiate themselves by providing a superior customer experience. Nowadays, there is no doubt that the key to success lies in the ability to better understand and act upon customers' requirements. The drastic increase of online information regarding issues, services and products has been accompanied by users' reviews. The amount of customer opinions has been increasing at a great speed, making the web more subjective and opinionated. The web surfers' comments cover almost all areas, as they are posted not only on specialized review sites, but also on most of the published news and blogs. Independent and unbiased consumers' reviews are known to be the most credible sources of product or service information and people tend to rely primarily on them when making a decision about a purchase. Little is known about how consumers' sentiments that are shared online, often referred to as 'e-sentiment,' influence products' market performance. Furthermore, we do not understand the comparative impact of e-sentiment as generated by consumers and traditional marketing actions such as advertising as controlled by marketers. When working with the web content concerning online reviews, blogs, forums etc, a performer deals with huge amounts of unstructured data in order to extract information. To be successful, he or she needs those data to be structured so that the necessary information becomes available. When extracted, the information needs to be aggregated and presented to the meaningful output in an understandable form. In the midst of the unstructured data collection process, information extraction or in this case sentiment extraction, aggregation of gathered information and presentation to the significant result, one is dealing with several innovative issues such as big data storing and analyzing large amounts of unstructured data, text mining deriving information from text and sentiment analysis finding out opinions from text.

According to Pew Research Center surveys in the U.S, nearly two-thirds of American adults (65%) use social networking sites, up from 7% when the center began systematically tracking social networking usage in 2005. 65% have used SNS (Social Networking Service) in 2015 (Figure. 1). An Internet service that strengthens personal relationships with acquaintances or creates new social networks to broaden one’s network of relationships, includes Minihompy, Blog, Twitter, Facebook, and Google etc.

(Units: %)



	2005	2006	2008	2009	2010	2011	2012	2013	2014	2015
Internet Users	10	16	34	50	60	65	67	73	74	76
All Adults	7	11	25	38	46	50	55	62	62	65

Source: Pew Research Center surveys

Fig. 1 SNS Usage Rate

In Figure 2, the SNS usage rates on each gender also are similar in both of males (64.7%) and females (65.1%). Meanwhile by age, the 20s (89.0%) have been the highest among all age groups, followed in order of the 30s (80.6%), the 40s (67.4%), etc in 2015. Especially, the SNS usage rates of 40s, 50s and 60s have been rapidly increasing from 2013 to 2015.

(Units: %)

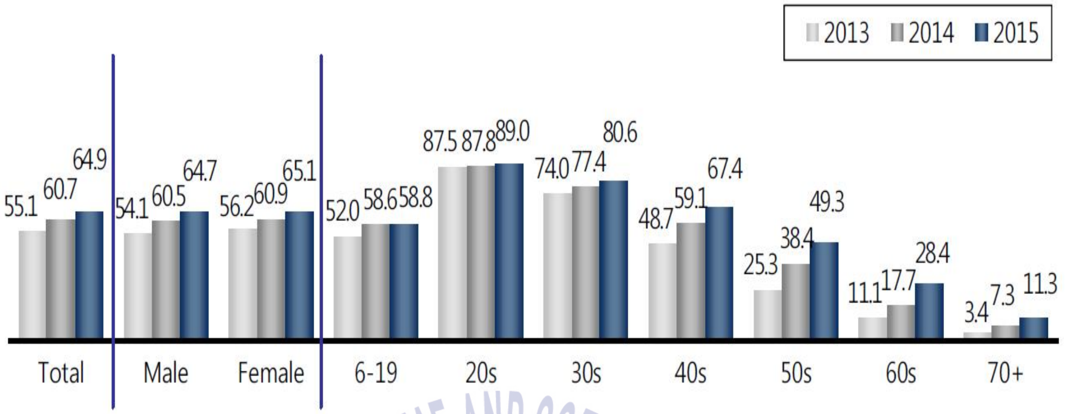


Fig. 2 SNS Usage Rate by Gender & Age

In addition, almost 9 out of 10 (88.4%) SNS users answered that they use a 'Profile-based service,' followed by 'Community (40.7%),' 'Blog (25.4%),' and 'Minihompy (15.4%),' etc. We pay attention to "Community," which is a group on the web interested in a specific issue, product or service (Figure. 3).

(Units: %)

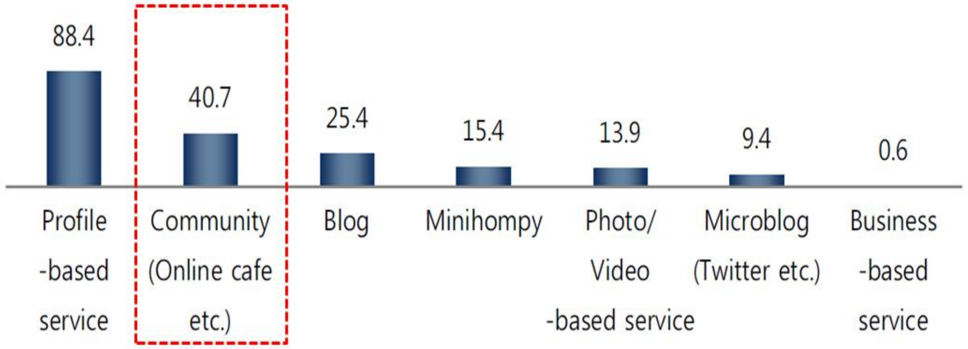


Fig. 3 SNS Usage by Service Type

We have recognized that the internet plays a role in not only conveying information, but the pathway diffusing the users' emotions. As the advent of various e-community, such as social networks or websites, is actualized, the interaction among e-community users contributes to the transition and diffusion of emotions (B. Kujawski et al, 2007). Customers regard the community as a resource to offer their opinions associated with products and services. Furthermore, it has been utilized as a medium that car companies can acquire customers' feedback. This research is promoted based on car e-community, which provides a space for communication with car companies and other customers who intend to sell and purchase cars in the near future. In this research, we first analyze the reviews of companies' vehicles on SNS and then deduct the users' demand characteristics. We collect the VOC (Voice of the customer) data from car e-community, extract the key words of vehicle and sentimental vocabularies by using text mining.

1.2 Necessity of Study

Leading companies build competitive strategies based on insights from VOC data. VOC is a term that describes customers' feedback about their experiences with products or services. The definition of VOC varies across authors. The definition of VOC originated with a paper in 1993 by Griffin and Hauser, who defined VOC as "a complete set of customer wants and needs; expressed in the customer's own language; organized the way the customer thinks about, uses and interacts with the product and service; and prioritized by the customer in terms of both importance and performance". According to Gerald M. Katz (2001), the VOC was first used to improve product development, though eventually referencing "any type of market research with customers". Thus, VOC data could be resources for business of organization (Figure. 4).

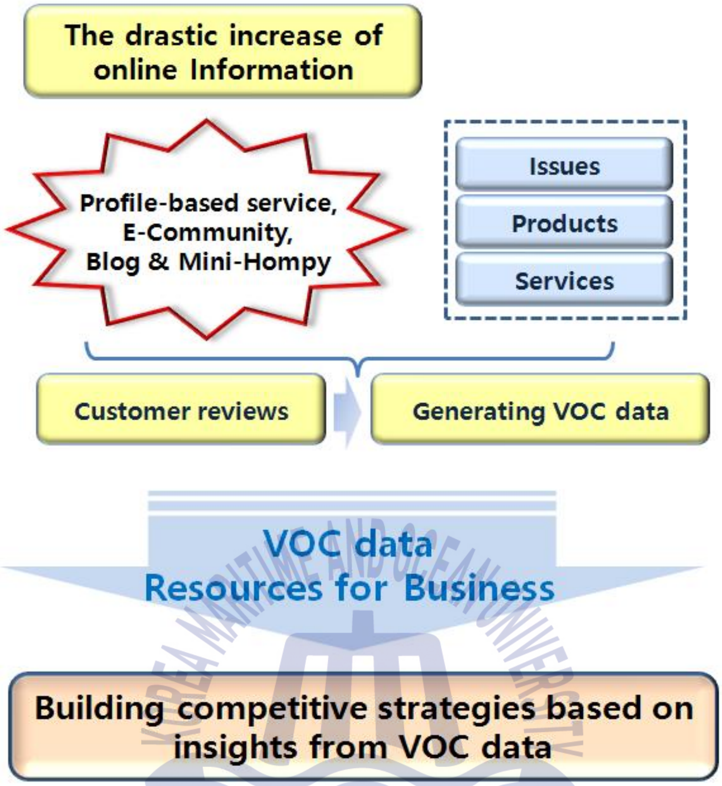

The drastic increase of online Information
Profile-based service, E-Community, Blog & Mini-Hompy
Issues
Products
Services
Customer reviews
Generating VOC data
VOC data
Resources for Business
Building competitive strategies based on insights from VOC data

Fig. 4 VOC : Resources for Business

Normally, vehicle companies discover problems through vehicle tests, inspection procedures, or information gathering. They may, for instance, review warranty claims or dealership service records, or consult consolidated insurance industry data. We believe, however, there are a lot of useful and hidden vehicle quality data embedded in social media that are largely untapped into by organizations. Recently, automotive companies like Chrysler have begun to employ "Twitter teams" to reply to whining tweets; but, detecting "whispers of useful information in a howling hurricane of noise" is a huge challenge and better filters are needed to extract meaning from the "blizzard of buzz" (Mentzas, 2011).

Consumer interest on products of companies generates VOC in negative and positive ways via a social media, and the information impacts on the performance of each company in the market. In this way, VOC has been recognized as one of the important factors in corporate performance by expanding the usage and boundary of VOC on the network. Nevertheless, the research in consideration of the VOC information of the product generated from social media is still lacking. In addition, the need to more efficiently analyze large amounts of unstructured data, review information of the product generated from social media by applying opinion mining techniques, has sustainedly emerged.

1.3 Purpose & Questions

The purpose of this study is to verify if the 7 sentiment domains; Sadness, Shame, Anger, Fear, Frustration, Delight and Satisfaction, extracted from social media are suitable for data for experiment through sentiment analysis, and investigate how sentiment of VOC has an influence on business performance. Specifically, we intend to investigate how VOC sentiment in online consumers' opinions on the products of vehicle manufacturing companies from car e-community, has an influence on the market share, which is one of the corporate performance indicators as the subject of e-community based on reliability of VOC data. Therefore, the specific questions dealt with in this study are as follows.

Are the data are reliable and suitable for experiment?

What is how to effectively construct sentiment lexicon and how do VOC sentiments have a significant influence on the market share?

1.4 Structure

This thesis consists of 5 chapters. Chapter 1, introduction is composed of back grounds of the study, necessity, purpose and questions.

In chapter 2, we inquired the precedent studies related to purpose of this research. The main contents of chapter 2 are listed in the order of importance of VOC, data mining, text mining and sentiment analysis and mentioned literature reviews of constructing sentiment lexicon. In chapter 3, based on literature reviews inquired in chapter, we depicted the research flow, and methodologies proposed.

This study consists of three phases. In phase I, we constructed the sentiment lexicon and 7 sentiment domains by analyzing sentiment of VOC on 26 auto companies. This composition is included in the chapter 4. Chapter 4 is the step of experiment and validation and composed of phase I, II and III.

In phase II, in order to retain the reliability of sentiment VOC data for experiment, we examined using the auto-correlation analysis and PCA (Principal Component Analysis). We analyzed how correlated they are through the auto-correlation analysis and which predictable information is included in the emotion. We validated whether 7 sentiment domains are appropriately classified into negativity and positivity by using PCA. Through this empirical study, we verified the reliability of sentimental data and would apply the data to the sustained study.

In phase III, we investigated how 7 domains impact on the performance of organizations, which is market share of four major auto companies by using linear regression analysis.

Finally, In chapter 5, we summarized the results from the experiment, provided managerial implication and limitation of the study, proposed directions for the future studies and advanced the research agenda.

2. Literature Reviews of VOC Analysis

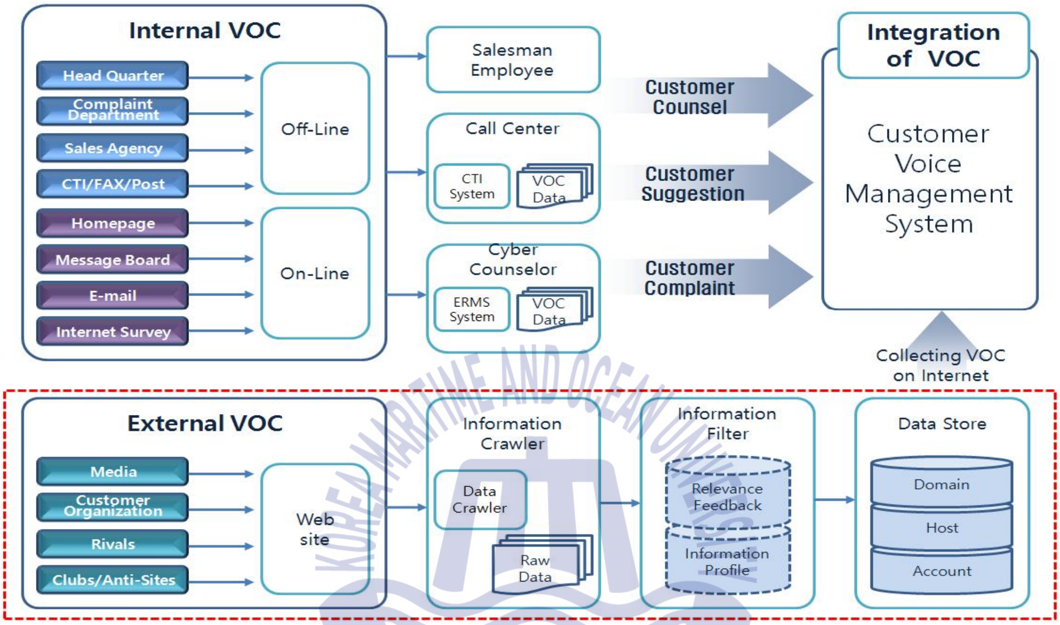
2.1 Importance of VOC

The VOC (Voice of the Customer) is a product development technique that produces a detailed set of customer wants and needs which are organized into a hierarchical structure, and then prioritized in terms of relative importance and satisfaction with current alternatives. The VOC process has important outputs and benefits for product developers. VOC provides;

- a delicate understanding of the customer's needs.
- a common, but useful language for the company carrying out.
- key input for the setting of appropriate specifications for the product and service.
- a practical medium for product and service innovation.

The purpose of VOC management in each organization is to specifically understand customers by figuring out whether customers are satisfied with the products and services provided by the company or not. It has built the structure of VOC system in order to grasp the changing needs and expectations of them from market. Figure 5 shows a integrated system structure for collecting VOC.

The VOC collection system is divided into two aspects; internal VOC and external VOC. In the internal VOC, there are two sources of links. One is off-line having four subordinate categories such as head quarter, complaint department, sales agency and kinds of means of communication, and the other utilizes on-line on internet. After the data are classified into types of customer counsel, suggestion and complaint, they are stored and utilized in accordance with the purpose in the CVMS (Customer Voice Management System).



source: Daumsoft

Fig. 5 Integrated VOC Collection System

On the other hand, the internal VOC composed of unstructured data from web site is collected by a information crawler tool as sorts of raw data. They are divided or categorized into relevant fields in the process of filtering and stored to types of domain, host and account. The data are managed depending upon the purpose in the CVMS. In this study, we focus on the research of the sentiment analysis on the external VOC because the study and practice of the internal VOC have been sufficiently proceed while those of the external VOC are lacking. Realizing the importance of VOC, organizations are now beginning to use them along with structured data to get insights that can be used to improve performance in terms of quality, operational efficiency, and revenue. However, They today resort to manual process to derive insights from voice of customers.

VOC analysis makes organizations be able to develop products and services to continuously reflect the needs of customers that frequently change by utilizing the VOC data. That is, by integrating VOC data collected from various routes, companies collect the changing needs of the customers in accordance with situation or local, analyze them, convert them into information required for the development of products and services and provide the solutions.

As reviewing the precedent studies related to VOC, Subramaniam et al. (2009) developed a system called BIVOC (Business Intelligence from Voice of Customer) where a significant portion of the VOC analysis and integration of VOC with structured information, but it has a limitation that it only dealt with the structured data. Although there were a study about how VOC with negative contents affect the interactions within an online brand community such as MyStarbucksIdea(Lee, Han & Suh, 2014), it just focused on the negative sentiment. On the other hand, there is also a study that intend to induce efficient use of VOC system by utilizing the concept recognition of VOC systems, satisfaction, and the recognition of influence and analyzing the differences in recognition between the customers and employees (Choi et al., 2011). Takeuchi et al. (2009) stressed out the necessity for business-oriented dialogue with our customers, beyond the marketing strategy, in order to improve the quality of products and services of the enterprises and the operational efficiency by using a text mining to analyze the telephone VOC data of the car rental help desk. it referred to the importance of the VOC.

However, the majority of the previous researches are dealing with the VOC importance and integrated system, even though we recognize that

VOC has a significantly large impact on customer satisfaction and corporate performance. Furthermore, a study of the contents of VOC data is even more sparse and lacking. That is why the actual VOC data have traits that it is difficult to be presented due to their attributes composed of complaints and protests. Lately, in addition, as internet and smart services are developing, although VOC warehouse a accumulate more VOC data, it is not easy to properly analyze and utilize because they consist of unstructured text data.

From this point of view, opinion mining, which is one field of text mining methods to analyze large amounts of VOC data, can be a useful alternative. Opinion Mining means a process of extracting, classifying, understanding and capitalizing the opinions that are exposed to the outside through a variety of contents such as online news and social media, and is carried out by utilizing a variety of techniques sentiment analysis (Liu, 2010). It is also easily able to be applied to determine the ranking to review data in order to enhance the search efficiency of customers' reviews for potential buyers in on-line shopping mall (Yune et al., 2010), and evaluate a movie by analyzing positivity and negativity after summarizing the movie reviews (Zhuang, 2006).

Furthermore, there is the intelligent investment decision-making model to take advantage of important information for stock price prediction by analyzing the news contents with opinion mining technique (Kim et al., 2012). Opinion mining is used to extract the opinion information that has been shown in the document. Specifically, it is used to classify the sentiment with dictionary such as WordNet, as using the linguistic rules such as the sentence structure, the relationship between the sentences, and pattern information of sentence elements, or classify the polarity and quantify the intensity of emotion with the natural language processing technology or machine learning. In particular, in the research for building linguistic resources, the sentiment lexicon is used, where positivity, negativity and neutrality are evaluated and tagged based on SentiWordNet. In this paper, we utilized WordNet and SentiWordNet to extract 7 sentiment domains.

2.2 Data mining

2.2.1 Concept & Functionalities

The term "data mining" is appropriately named as "Knowledge mining from data" or "Knowledge mining". This technique is actually called as a data mining or Knowledge Hub or simply KDD (Knowledge Discovery Process). It is the non-trivial extraction of implicit, previously unknown and potentially useful information from data in data warehouse. Data mining is the process of analyzing data from different perspectives and summarizing them into useful information that can be used to increase revenue, cuts costs, or both. Data mining software allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. As the data are available in the different formats in order to be taken the proper action. Not only to analyze these data but also make a good decision and maintain the data. As and when the customer will require the data should be retrieved from the database and make the better decision. The important reason that attracted a great deal of attention in information technology the discovery of useful information from large collections of data industry towards field of data mining is due to the perception of "we are data rich, but information poor". There is huge volume of data but we hardly able to turn them in to useful information and knowledge for managerial decision making in business. In order to generate information it requires massive collection of data. It may be a variety of formats like audio, video, numbers, text, figures, and hypertext formats. To completely utilize data; the data retrieval is not enough, it requires a tool for automatic summarization of data, extraction of the essence of information, and the discovery of patterns in raw data. With the large amount of data stored in files, databases, and other repositories, it is increasingly important, to develop powerful tool for analysis and interpretation of such data and for the extraction of interesting knowledge that could help in decision-making. The only answer to all above is "Data Mining". Data mining is the extraction of hidden predictive information from large databases; it is a powerful technology with great potential to help organizations to focus on the most crucial information in their data warehouses (Larose, 2005).

Data mining deals with what kind of patterns can be mined. On the basis of kind of data to be mined there are two kinds of functions involved in data mining. One is "Descriptive minning", the other is "Classification and Prediction."

Descriptive mining tasks characterize the general properties of data in database. Predictive mining tasks perform inference on the current data to make predictions. Descriptive functions are class and concept description, mining of frequent pattern, associations, correlations and clusters.

In classification and prediction, predictive data mining tasks practice inference of the current data in order to make prediction. Predictive data mining involves using some variables or fields in the data set to predict unknown or future values of other variables of interest, and produces the model of the system described by the given data set. The goal of predictive data mining is to create a model that can be used to perform tasks such as classification, prediction or estimation. The goal of a predictive data mining model is to forecast the future outcomes based on the past records with known answers. Classification requires the data mining algorithm to divide the input space in such a way as to separate the examples based on their class.

The data mining project consists of six phases (Larose, 2005 ; Chapman et al., 2000). The sequence of the phases is flexible, not rigid. Moving back and forth between different phases is always adopted. It relies on the outcome of each phase. There are six main phases.

1) Business understanding

This initial phase concentrates on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition, and a preliminary goal designed to accomplish the objectives.

2) Data mining understanding

The data understanding phase starts with data collection and proceeds with activities to become familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information.

3) Data preparation

This phase covers all activities to construct the final dataset. Data preparation tasks are likely to be performed more times, and not in any appointed order. Tasks include table, record, attribute selection as well as transformation and cleaning of data for modeling tools.

4) Modeling

In modeling phase, a variety of modeling techniques are selected and applied their parameters are measured to optimal values. Several techniques for the same data mining problem type exist. Some techniques have specific requirements on the form of data. Therefore, it is often necessary to step back to the data preparation phase.

5) Evaluation

At this phase, the model (or models) built appears to obtain high quality from a data analysis perspective. Before promoting to final deployment of the model, it is crucial to evaluate the model more thoroughly, and review the steps performed to build the model, to be certain it appropriately achieves the business objectives. A key objective is to determine if there is some important business issue that has not been considered sufficiently. At the end of this phase, a decision on the use of the data mining results should be reached.

6) Deployment

The goal of the model is to increase knowledge from the data, the knowledge gained will have to be organized and presented in a way that the customer can use it. Construction of the new model is not the end of the project in general. Even though the goal of the model is to increase knowledge from the data, the knowledge gained will need to be organized and presented in a way that the client can use. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process.

Data mining functionalities include classification, clustering, association analysis, time series analysis, and other analysis.

Classification is the process of finding a set of models or functions that depict and distinguish data classes or concepts, for the purpose of predicting the class of objects whose class label is unknown. Classification is important for decision making management. Given an object, assigning it to one of predefined target categories or classes is called classification. The purpose of classification is to accurately forecast the target class for each case in the data (Kesavaraj, 2013). For example, a classification model could be utilized to identify loan applicants as low, medium, or high credit risks (Song, 2011). There are various methods to classify the data, including decision tree induction, frame-based or rule-based expert systems, hierarchical classification, neural networks, Bayesian network, and support vector machines

Clustering analyzes data objects without consulting a known class model. Clustering algorithms (Jain et al., 1988) divide data into meaningful groups so that patterns in the same group are similar in some sense and patterns in different group are dissimilar in the same sense. Searching for clusters involves unsupervised learning (Ansari et al., 2013). In information retrieval, for example, the search engine clusters billions of web pages into different groups, such as news, reviews, videos, and audios. One straightforward example of clustering problem is to divide points into different groups (Song, 2011).

Association analysis is the discovery of association rules displaying attribute-value conditions that frequently occur together in a given set of data. Association rule mining (Agrawal et al., 1993) focuses on the market basket analysis or transaction data analysis, and it targets discovery of rules showing attribute value associations that occur frequently and also help in the generation of more general and qualitative knowledge which in turn helps in decision making (Gosain & Bhugra, 2013).

Time series analysis comprises methods and techniques for analyzing time series data in order to extract meaningful statistics and other features of the data. A time series is a collection of temporal data objects; the features of time series data include large data size, high dimensionality, and updating continuously. In general, time series task relies on 3 parts of components, including representation, similarity measures, and indexing (Fu, 2011; Esling, 2012).

2.2.2 Methodologies of Data mining

Decision Trees are like those applied to decision analysis where each non-terminal node represents a test or decision on the data item considered. Depending on the outcome of the test, one chooses a certain branch. As in data mining applications, very large training sets with several million examples are common; a decision tree classifier scales well and can handle training data of this magnitude. So, for classifying large data sets, our focus mainly on decision tree classifiers. Tree-shaped structure that represent sets of decisions. In decision tree node represent a test on an attribute value, branch represents an outcome of the test and tree leaves represent classes or class distribution. Decision trees partition the input space into cells where each cell belongs to one class. The partitioning is represented the sequence of tests. Interior node in the decision tree tests the value of some input variable, and the branches from the node are labelled with the possible results of the test. The leaf nodes represent the cells and specify the class to return if that leaf node is reached. The classification of a input instance is start at the root node and, depending on the results of the tests, following the appropriate branches until a leaf node is reached (Rokach, 2008). Decision tree is represented in figure 6. A disadvantage of DT is that trees use up data very rapidly in the training process. They should never be used with small data sets. They are also highly sensitive to noise in the data, and they try to fit the data exactly, which is referred to as "over-fitting."

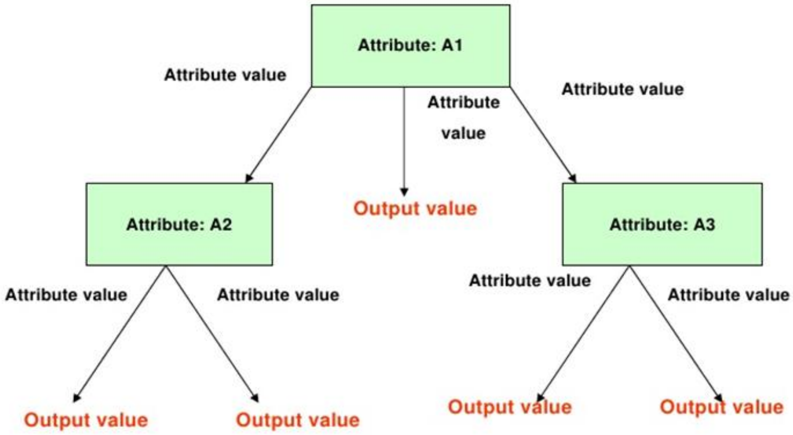


Fig. 6 An Example of Decision Trees

The next methodology is "Neural networks". In more practical terms neural networks are non-linear statistical data modeling tools. They can be used to model complex relationships between inputs and outputs or to find patterns from data. Utilizing neural networks as a tool, data warehousing organizations are extracting information from data sets in the process, data mining. The difference between the data warehouses and ordinary databases is that there is actual manipulation and cross fertilization of the data helping users makes more informed decisions. Neural networks fundamentally comprise three pieces: the architecture or model; the learning algorithm; and the activation functions. Data mining technique such as neural networks is able to model the relationships that exist in data collections and can therefore be used for increasing business intelligence across a variety of business applications (Andrews, 1995).

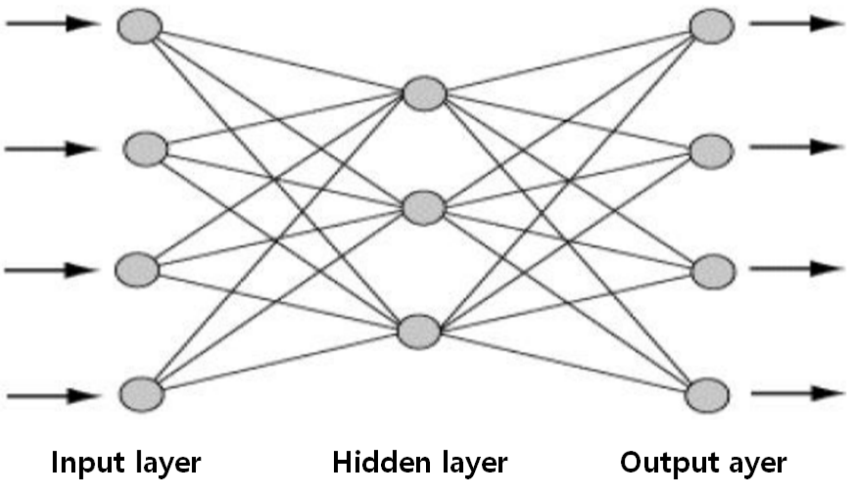


Fig. 7 Neural Networks

It is shown in figure 7. Artificial neural network have become a useful tool in tasks like pattern recognition, decision making or forecasting applications. It is one of the newest signals processing method. ANN is an adaptive, non linear system that learns to practice a function from data and that adaptive phase is normally training phase in which system parameter is change during operations. After the training is complete, the parameters are fixed. If there are a lot of data and problem is poorly understandable, then utilizing ANN model is accurate, the non linear traits of ANN provide it a lot of flexibility to achieve input output map.

The next methodology is "GA (Genetic algorithm)." GA attempts to incorporate ideas of natural evaluation. The general idea behind GAs is that we can build a better solution if we somehow combine the "good" parts of other solutions through schemata theory, just like nature does by combining the DNA of living beings (Agarwal, 2012). The GA was developed by John Holland in 1970. They are based on the genetic processes of biological organisms. As many generations go over, populations in nature evolve according to the principles of natural selection and "survival of the fittest," first definitely stated by Charles Darwin in the Origin of Species. GAs are adaptive methods which may be used to solve search and optimization problems. After a great deal of new generations constructed with the help of the described mechanisms, one obtains a solution that cannot be improved any further. This solution is taken as a final one (Holland, 1975). When GAs are used to solve problem, the solution has three distinct stages.

1) Selection: The selection function chooses the parents using roulette wheel and uniform sampling, based on expectation and number of parents.

2) Crossover: The crossover function is position independent. This crossover function creates the crossover children of the given population using the available parent.

3) Mutation: Produce new gene individuals by recombining features of their parents.

The next methodology is "Rule Extraction." Andrews (1995) identifies three categories for rule extraction procedures: decompositional, pedagogical, and eclectic.

Decompositional rule extraction includes the extraction of rules from a network in a neuron-by-neuron series of steps (Darrah et al., 2005). The drawbacks of decompositional extractions are time and computational limitations. The advantages of decompositional techniques are that they seem to offer the prospect of generating a complete set of rules for the neural network.

Pedagogical rule extraction (Nayak et. al., 1997) deals with the entire network as a black box. In this approach, inputs and outputs are matched to each other. The decompositional approaches can generate intermediary rules defined for internal connections of a network possibly between the input layer and the first hidden layer.

The eclectic approach is the use of those techniques that incorporate some of a decompositional approach with some of a pedagogical approach. There are several main rule formats. Rule extraction algorithms will generate rules of either conjunctive form or subset selection form, commonly referred to as M-of-N rules named for the primary rule extraction that takes advantage of the form. All rules follow the natural language syntactical if-then prepositional form.

2.3 Text Mining

Text mining, called text analysis, text analytics, text data mining, automatic text analysis, and computer-based text analysis, is the analysis of data composed of texts in order to discover hidden patterns, features, and relationships. Beyond the linguistic and content analysis of texts, text mining often includes extended functions, such as statistical analysis, database building, and outcome visualization. In spite of the different terminologies and technologies used, the general goal of text mining is to convert the unstructured text information into structured information, which can be analyzed with traditional data mining and statistical techniques. Generally, text analysis has been utilized in other disciplines such as psychology to analyze the content of written information and to predict psychological states and behaviors of individuals (Bantum & Owen, 2009). Some initial evidence regarding the validity and utility of text mining methods has been provided in psychology (Alpers et al., 2005; Kahn et al., 2007). However, since most of the psychological studies applying text mining were carried out in the context of therapeutic discourse and health-related disorders such as sentimental writing of trauma and cancer narratives, which involved intense emotions, it is unclear whether the text mining method could be safely employed to capture VOC information in consumers' normal lives. Text mining can be conducted from a qualitative approach, a quantitative approach, or a combination of both. Qualitative analysis generates non-numerical information and is generally used to understand the reasons and motivations from the bottom line, to identify the underlying themes or relationships, and to develop hypotheses (Kozinets, 2002). In contrast, the quantitative approach regards texts as objective data and generates numerical data. This way of approach is more extensively utilized (e.g., Liu, 2006; Srinivasan, 2012) principally because its output can be directly utilized by convention of a statistical analysis. Typical text mining tasks include information extraction, text categorization, text clustering, document summarization, and association analysis (e.g., Pennebaker et al., 2003; Gupta & Lehal, 2009; Ramanathan & Meyyappan, 2013). Information extraction is the most basic function of text mining. This technique can be used to identify the important market terms, issues and themes such as brand names and product traits from social media. Categorization is applied to classify and assign texts into predefined categories or subjects. In categorization, computer programs often deal with a text as a basket of words and count the word frequencies. For example, the words "trouble" and "anger" might be assigned to the category of "negative emotions." This approach is widely used to identify communicators' attitudes or emotional states in sentiment analysis. Clustering is a technique used to group similar documents. It is different from categorization in that it does not use pre-defined categories, but instead clusters documents in real time. Association analysis is to find associations for a given term based on counting co-occurrence frequencies. For instance, if one brand always appears in social media with positive adjectives, it indicates that consumers may get a positive image from that brand. Summarization is to summarize the important concepts in texts while reducing the length and detail of a document. Based on a big collection of text materials, summarization can also help identify the market trends, such as the changes of consumer preferences as time goes by. In addition, text mining tools often have other additional functions such as visualization, in which a graphical information representation can be developed and displayed. A variety of computer-based text mining tools have been developed with the advances in computer technology, which makes text mining easier. Recently, a couple of innovative studies have applied text mining techniques to acquire essential information from VOC. For example, Aggarwal et al. (2009) used lexical semantic analysis, a text mining approach, to evaluate brand positions relative to competitors. They, first of all, collected web content related to their target-brands through Google. They then carried out lexical semantic analysis to examine the co-occurrence of brand names and key adjective descriptors. if Samsung frequently appears on web pages with the adjective "reliable," it indicates that "reliable" is a main feature of Samsung.

2.4 Sentiment Analysis

Sentiment analysis, also known as opinion mining, refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. A basic task in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect level-whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral. Advanced, "beyond polarity" sentiment classification looks, for instance, at emotional states such as "angry," "sad," and "happy." Sentiment analysis is traditionally used to classify the positive or negative emotions composed of text comments and extract an object representing the writer's opinion (Zhuang et al, 2007; Lerner at al, 2000). The study of sentiment analysis has been started from the late 1990s has been fixed in the increasingly important field of research since 2000s (Thelwall, 2011; Argamon, 1998; Kessler & Spertus, 1997; Li, 2010). Studies, which has been progressing until now, mainly has focused on the classification of positive or negative opinions such as movie and product purchase reviews (Zhuang, 2006; Pang, 2002; Whitelaw, 2005). However, recent studies tend to regard feeling as precisely subdivided units overcoming a simple classification of it (Bollen et al., 2011). Thus, these classification methods play a crucial role in providing the research domain of sentiment analysis with deeper understandings and suggestions (Wilson, 2009; Esuli, 2006). We classified emotions into 7 domains (Sadness, Shame, Anger, Fear, Frustration, Delight and Satisfaction) by using SentiWordNet.

Table 1 Emotion Domain & Basic Sentiment Words

Domain	Basic Sentiment Words
Sadness	*sadness, loneliness, unhappiness, depression*
Shame	*shame, guilt, regret, embarrassment*
Anger	*anger, irritation, disgust, rage*
Fear	*fear, worry, anxiety, nervousness*
Frustration	*frustration, resignation, powerlessness, despair*
Delight	*delight, pleasure, joy, happiness*
Satisfaction	*Satisfaction, gratification, fulfillment*

Table 1 shows negative and positive emotion domain and subdivided basic sentiment words, which utilize a standard for classification of emotion (Diener, 2006; Tronvoll, 2011).

In the opinion mining literature, it is presumed that heavily negative and positive reviews will be indicative of advantages and disadvantages of the product and service. However, whether this presumption is true for automotive sales has not been tested in prior research. There is some evidence that generic sentiment analysis fails when applied across domains. Loughran and McDonald (2011) found that sentiment-indicative words differ across domains: specifically, in the field of finance, sentiment indicators were different from sentiment marker words previously thought to be generally applicable to all fields. O'Leary (2011) found that generic positive and negative dictionaries had some limitations in describing negative behavior in the stock market, and suggested that domain specific terms be accounted for to improve the quality of the analysis. In the vehicle industry, therefore, generic sentiment polarity analysis may be insufficient. A thread poster may be more aggrieved by a malfunctioning air conditioner than with a sticky accelerator pedal, yet the sticky pedal is almost certainly a more serious defect. For instance, to enable proper investigation, the defect must be associated with the troublesome component, so hazard analysis can be performed (Jesty et al., 2000; Ward et. al, 2009). Table 2 summarizes previous research on the organizational use of text analysis of traditional internet and social media, for competitive intelligence, in various application domains. For each study, table 2 shows the medium, domain, and competitive intelligence perspective., for the study. We classified competitive intelligence perspectives using Vedder et al. (1999). Table 2 highlights the research gap, which we aim to address in this paper: the application of text mining to vehicle market analysis in the on-line car community.

Table 2 Text Analysis Studies via Social Media.

Study	Medium	Domain	Competitive intelligence perspective
Lee, Han & Suh, 2014	e-community	Customer reviews	Customer
Kim & Jin, 2013	SNS	Customer reviews	Customer
Kim et al., 2012	News	Stock	Market
Li & Wu, 2010	Blog	Sports	Market
Coussement & Poel, 2008	Email	Customer complaints	Customer
Spangler & Kreulen, 2008	Email	Customer complaints	Customer
Romano et al., 2003	Product reviews	Movie	Customer
Duan, Gu, & Whinston, 2008	Product reviews	Movie	Product
Schumaker & Chen, 2009	News	Stock	Market
Tetlock et al., 2008	News	Stock	Market
Finch, 1999	News	Power tools	Product

As the precedent studies on construction of sentiment lexicon, Turney and Littman (2002) constructed semantic orientation from a hundred-billion words using corpus. Lately, in addition to the research predicting the direction of the stock Index by utilizing a domain-specific sentiment dictionary (Yu et al., 2013), An and Kim (2015) built a Korean sentiment lexicon by using collective intelligence, and Jo and Choi (2015) established sentiment lexicon based on OAR (Opinion Antonym Rule) algorithm. Besides, the study using LP (Label Propagation) extract the sensibility dictionary through the proximity between the words. In particular, Song and Lee (2013) proposed a study, which verified that the accuracy of sentiment analysis had improved when using a specialized lexicon rather than a generic sentiment dictionary. Although this study is promoted based on the generic lexicon, we constructed the sentiment lexicon by extracting the terminologies related to vehicle in order to compensate the defect of generic lexicon.

2.5 Research Trends in Korea

There are various studies that have conducted opinion mining in Korean. Machine learning and transformation-based learning method, as two types of Korean sentiment analysis techniques, have been conducted (Yang & Ko, 2014). As researches applying them, analysis of public opinion on recent trends through Twitter and forecasting the stock market by using sentiment on news (Lee & Lee, 2013; Kim et al., 2014). Kim et al. (2012) constructed an evaluation set that can measure the sentiment analysis. It is able to be applied to measuring not only the polarity of the user's sentiment but also the type and intensity of sentiment. In addition, after detailed evaluation elements for a restaurant such as taste, service, atmosphere, price, food, sheep, sanitation, parking, and representative menu were selected, sentences including evaluation elements-opinion linkage were extracted from restaurant reviews (Park et al., 2013). Besides, sentiment analysis is simply applied to determining the ranking of the product review data in order to increase the efficiency for potential buyers to retrieve product reviews in the online shopping mall (Yune et al., 2010). Furthermore, Chae et al. (2012) proposed a new model of crisis management, which is also able to recognize the crisis of corporate and extract key information on risk by early detecting opinions on social media.

The key factors of opinion mining for Korean are morphological analysis and construction of sentiment lexicon. A morpheme is the minimum unit at the morphological level of language. Morphological analysis can be defined as assigning categories to each morpheme after dividing the word of a sentence into morpheme units (Shim & Yang, 2004). The morphological analysis of the data in Korean has been carried out by "Kokoma morpheme analyzer" developed by IDS laboratory of Seoul National University in general. In order to classify the polarity of the separated morphemes through morphological analysis, sentiment lexicon that evaluates positivity / negativity / neutrality in the meaning level of words or words in individual language contents is utilized (Yu et al., 2013). Song and Lee (2011) verified that using sentiment lexicon based on the characteristics of domain improves the accuracy of emotion evaluation, and suggested the ways to construct and use specialized sentiment lexicon based on this research.

However, although the study of Korean sentiment analysis has been continuously proceeded since the early 2000s, it is lacking as ever due to inadequate conditions such as intricate structure of Korean morpheme and the lack of programs or tools related to Korean sentiment analysis. In particular, not frequency, there is a difficulty in scoring for the polarity or intensity of sentiment, Therefore, a variety of research methodologies and techniques on Korean sentiment analysis, which are suitable for Hangul system, are urgently needed and should be developed. The future research will reflect these problems.

3. Methodology

3.1 Research Flow

In this study, we extract the key words of emotions in VOC data that had been posted on car-related on-line community from 2013 to 2015, and intend to analyze the correlation between negative and positive key words and contribution to market share. This study is performed following the research flowchart as figure 8. Schematic contents are as follows. We build the database, crawling customers' opinion information from the car-related online community and identify the part of speech (POS) information about words in the customers' opinion by using a POS tagging function provided by TextAnalysisOnline. We construct the negative and positive emotional vocabulary group; that is, sentimental lexicon. Based on the data, we measure the polarity. As we investigate the previous studies, regardless of the domestic or international, there are only a few empirical analysis about the correlation between the data on consumers' sentiment and the market on social networks because it is difficult to extract the key factors with sentiment and prediction associated with each industry. Therefore, we perform the empirical analysis on the 7 sentiment domains that have already been extracted via the previous research as a measurement tool. In case of literature research, while a few research analyzed the opinion such as customer complaints, prediction of stock index, and sales and image of the product, taking advantage of the data, in the present study, we grasped and analyzed how correlated they are through the correlation analysis and which information is included in the 7 sentiment domains through auto-correlation analysis and principal component analysis.

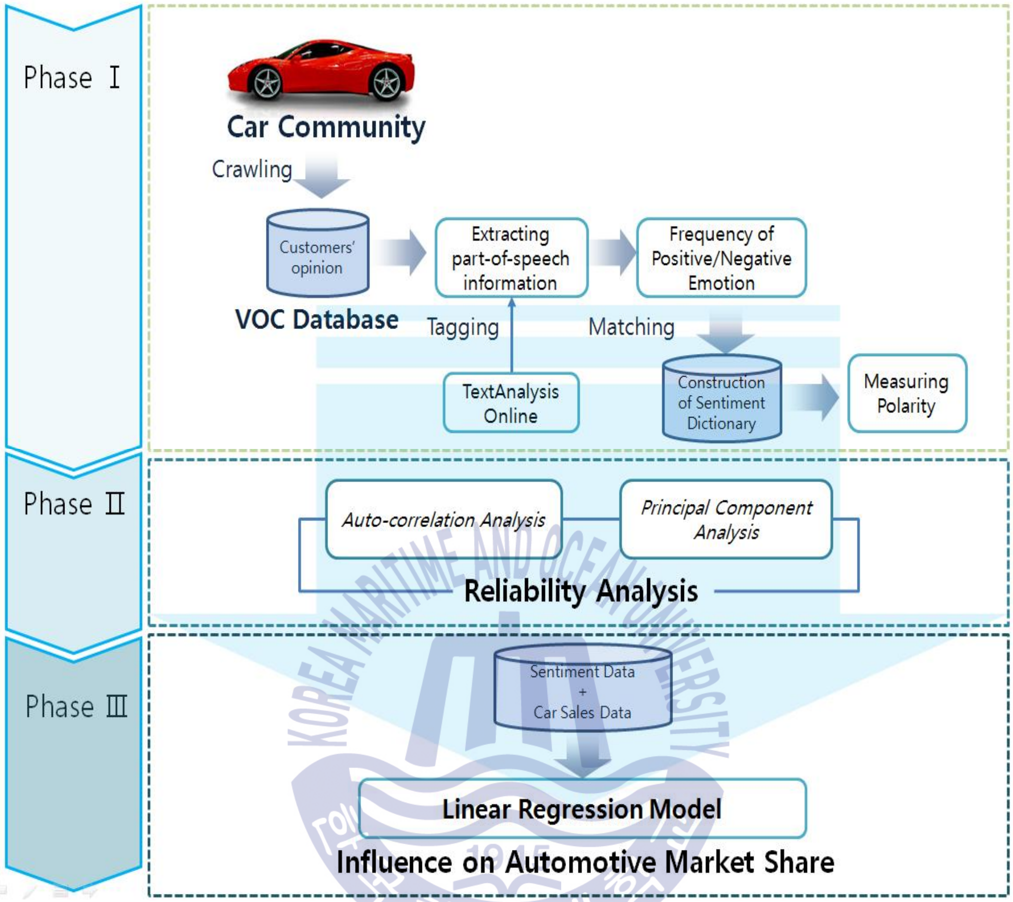

Phase Ⅰ
Car Community
Crawling
Customers' opinion
VOC Database
Extracting part-of-speech information
Frequency of Positive/Negative Emotion
Tagging
Matching
TextAnalysis Online
Construction of Sentiment Dictionary
Measuring Polarity
Phase Ⅱ
Auto-correlation Analysis
Principal Component Analysis
Reliability Analysis
Phase Ⅲ
Sentiment Data + Car Sales Data
Linear Regression Model
Influence on Automotive Market Share

Fig. 8 Research Flow

Specifically, at first, we grasped auto-correlation of 7 sentiment domains that contain sales information, and checked whether it contains information that can predict the future. At this time, the 7 domains which are the main variables are composed of a combination of words that contains the feeling of the car market. In other words, after collecting information from the car community, we removed the buzz data, such as unwanted spam and noise and classified 7 sentiment domains (Sadness, Shame, Anger, Fear, Frustration, Delight and Satisfaction) of the auto sales market. It is possible to forecast the actual market share, if they have the auto-correlation of sentiment, because it contains information of sales in the emotion itself. Secondly, we checked which main factors 7 domains have by using PCA. We precisely analyzed the auto-correlation of each emotion and proposed a methodology to test the reliability of the information of the car community. In Phase III, based on the data reliability, we built the two kinds of linear regression model and verified how the sentiments such as positivity, negativity and neutrality have an influence on the performance of organizations, which is auto market share in this paper.

3.2 Proposed Methodologies

3.2.1 Sentiment Analysis

Sentiment analysis also known as opinion mining refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. Sentiment analysis is widely applied to reviews and social media for a variety of applications, ranging from marketing to customer service.

In general, sentiment analysis focuses on determining the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document. The attitude may be his or her judgment or evaluation based on appraisal theory, which is a theory in psychology that emotions are extracted from our evaluations of events such as appraisals or estimates that cause specific reactions in different people, affective state, that is, the emotional state of the author as writing, or the intended emotional communication, that is to say, the emotional effect the author wishes to have on the reader.

Sentiment analysis consists of three steps in total. The first phase is the "Data Collection", which collects information from various social media. The second is the "Subjectivity Detection," which filters out only the parts of the user's subject that are revealed from the holistically collected information. In the third stage, "Polarity Detection" is performed, which is a process of classifying the extracted emotion data into two extremes of 'like' and 'dislike'.

In the step of Data Collection, the vast amount of data on the Internet is a key element for sentiment analysis. Data can be crawled from not only public data sources such as blogs, bulletin boards, product evaluations, but also personal social network sites such as Twitter and Facebook. The search engine is usually used to collect data to apply sentiment analysis techniques. The search engine receives the user's query and collects all the documents including the query, ie, related data including various stars, reviews, and comments. However, since the data collected includes some parts not to be related to user's emotion, it is necessary to perform "Subjective Detection" which extracts only the part to be subjected to emotional analysis after data collection. A Web crawler is an internet bot which systematically browses the World Wide Web, typically for the purpose of Web indexing (web spidering). Web search engines and some other sites use Web crawling or spidering software to update their web content or indices of others sites' web content. Web crawlers can copy all the pages they visit for later processing by a search engine which indexes the downloaded pages so the users can search much more efficiently.

Crawlers consume resources on the systems they visit and often visit sites without tacit approval. As the number of pages on the internet is extremely large, even the largest crawlers fall short of making a complete index. For that reason search engines were bad at giving relevant search results in the early years of the World Wide Web, before the year 2000. This is improved greatly by modern search engines, nowadays very good results are given instantly. Crawlers can validate hyperlinks and HTML code. They can also be used for web scraping (Wikipedia).

In the process of "Subjectivity Detection," after collecting the necessary text, it is necessary to separate and classify only the text elements to be used for sentiment analysis. In general, the text collected on the web excludes parts of the sentence which are not relevant to emotion. In addition, the process of filtering out personal information such as the name and sex of the text author is used to avoid problems in collecting excess information. For example, "Text Analytics" is an information analytics technology that extracts meaningful information from large amounts of unstructured text. Text analytics divides the components of a given text into three areas: positivity, negativity, and neutrality. Sentence that only states a fact like ′I bought a new smart phone today.′ is excluded from the analysis because it is classified into neutrality that can not judge a value at all.

In the "Polarity Detection" step, "Polarity Detection" operation is performed to determine whether the given data is ′positive′ or ′negative′. The computer detects positive and negative words in the text, quantifies them, and applies statistical techniques. For example, after assigning scores or weights according to ′frequency,′ ′affinity′ such as affirmation or negation of each word in a document, and then summing or averaging the scores indicated by each word, whether the entire text is positive or negative is determined. Sentiment analysis includes polarity analysis of ′document,′ ′attribute′ unit, and analysis technique using lexicon. The polarity analysis of document unit mainly uses machine learning, which involves two processes, ′training′ and ′classification.′ Machine learning is a field of artificial intelligence that makes a computer learn to formulate a specific pattern that exists in a given piece of data, and allow the computer to interpret other similar data. Models that mimic the human learning process allow the computer to interpret new data or predict the near future, such as weather phenomena or rise and fall of stock prices. In the training phase of machine learning, the user manually extracts specific values from documents classified as positive or negative and generates 'training data.' The computer then "classifies" whether the entire document is positive or negative through a certain machine learning model.

3.2.2 Auto-correlation Analysis

Auto-correlation is differently defined in a variety of fields of study, and not all of these definitions are equivalent. In some fields, the term is used interchangeably with autocovariance. In statistics, the auto-correlation of a random process is the correlation between values of the process at different times, as a function of the two times or of the time lag. Let X be a stochastic process, and t be any point in time. (t may be an integer for a discrete-time process or a real number for a continuous-time process.) Then X_t is the value (or realization) produced by a given run of the process at time t. Suppose that the process has mean μ_t and variance σ_t^2 at time t, for each t. Then the definition of the auto-correlation between times s and t is

$$R(s,t) = \frac{E[(X_t - \mu_t)(X_s - \mu_s)]}{\sigma_t \sigma_s},$$

where "E" is the expected value operator. Note that this expression is not well-defined for all-time series or processes, because the mean may not exist, or the variance may be zero (for a constant process) or infinite (for processes with distribution lacking well-behaved moments, such as certain types of power law). If the function R is well-defined, its value must lie in the range [−1, 1], with 1 indicating perfect correlation and −1 indicating perfect anticorrelation.

If X_t is a wide-sense stationary process then the mean μ and the variance

σ^2 are time-independent, and further the auto-correlation depends only on the lag between t and s: the correlation depends only on the time-distance between the pair of values but not on their position in time. This further implies that the auto-correlation can be expressed as a function of the time-lag, and that this would be an even function of the lag $\tau = s - t$. This gives the more familiar form

$$R(\tau) = \frac{E[(X_t - \mu)(X_{t+\tau} - \mu)]}{\sigma^2},$$

and the fact that this is an even function can be stated as

$$R(\tau) = R(-\tau).$$

It is common practice in some disciplines, other than statistics and time series analysis, to drop the normalization by σ^2 and use the term "auto-correlation" interchangeably with "autocovariance." However, the normalization is important both because the interpretation of the auto-correlation as a correlation provides a scale-free measure of the strength of statistical dependence, and because the normalization has an effect on the statistical properties of the estimated auto-correlations (wikipedia).

3.2.3 Principal Component Analysis (PCA)

Principal component analysis (PCA) is a statistical procedure that uses an to convert a set of observations of possibly correlated variables into a set of values of variables called principal components. The number of principal components is less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has the largest possible (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is to the preceding components. The resulting vectors are an uncorrelated orthogonal basis set. PCA is sensitive to the relative scaling of the original variables (Pearson, K, 1901). PCA is mostly used as a tool in exploratory data analysis and for making predictive models. PCA can be done by eigenvalue of a data covariance (or correlation) matrix or singular value decomposition of a data matrix, usually after mean centering (and normalizing or using Z-scores) the data matrix for each attribute (Abdi & Williams, 2010). The results of a PCA are usually discussed in terms of component scores, sometimes called factor scores (the transformed variable values corresponding to a particular data point), and loadings (the weight by which each standardized original variable should be multiplied to get the component score) (Shaw, 2003). In fact, PCA is performed to extract the component, which is representative of the variables. The entire process of PCA is as follows.

Assuming that the overall average of the dataset is 0, (if not, subtract the mean from the dataset) the principal component W_1 of dataset X is defined as follows:

$$W_1 = arg \max_{\|W\|=1} E\{(W^T X)^2\}$$

"arg max" indicates X, which makes the value of function f(x) maximize.

As k-1 principal components are already given, the kth principal component can be found by subtracting the previous k-1 principal components:

$$\hat{X}_{k-1} = X - \sum_{i=1}^{k-1} W_i W_i^T X ,$$

and the next principal component is newly found after subtracting this value from the data set.

$$W_k = \arg \max_{\|W\|=1} E\{(W^T \hat{X}_{k-1})^2\}.$$

Thus, after Karhunen-Loève transform finds the singular value decomposition of the data matrix X,

$$X = W\Sigma V^T,$$

this is equivalent to finding a partial data set Y by mapping X to a partial space defined by L singular vectors, W_L.

$$Y = W_L^T X = \Sigma_L V_L^T$$

The singular value vector matrix W of X is equal to the eigenvector matrix W of the covariance $C = XX^T$.

$$XX^T = W\Sigma^2 W^T$$

The eigenvector with the largest eigenvalue corresponds to the dimension with the strongest correlation in the data set.

3.2.4 Linear Regression

Linear regression is an approach for modeling the relationship between a scalar dependent variable Y and one or more explanatory variables (or independent variables) denoted X. The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression (Freedman, 2009). This term should be distinguished from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable (Alvin et al., 2012).

In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data.

Such models are called linear models (Seal, 1967). Most commonly, the conditional mean of Y given the value of X is assumed to be an affine function of X; less commonly, the median or some other quantile of the conditional distribution of Y given X is expressed as a linear function of X. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of Y given X, rather than on the joint probability distribution of Y and X, which is the domain of multivariate analysis.

Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications (Yan & Xin, 2009). This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine. Linear regression has many practical uses. Most applications fall into one of the following two broad categories:

If the goal is prediction, or forecasting, or error reduction, linear regression can be used to fit a predictive model to an observed data set of Y and X values. After developing such a model, if an additional value of X is then given without its accompanying value of Y, the fitted model can be used to make a prediction of the value of Y.

Given a variable y and a number of variables X_1, ..., X_p that may be related to Y, linear regression analysis can be applied to quantify the strength of the relationship between y and the X_j, to assess which X_j may have no relationship with Y at all, and to identify which subsets of the X_j contain redundant information about Y.

In this paper, linear regression analysis was applied to quantify the strength of the relationship between VOC sentiment and automotive market share.

4. Experiment & Analysis

4.1 Phase I : Constructing Sentiment Lexicon & 7 Sentiment Domains

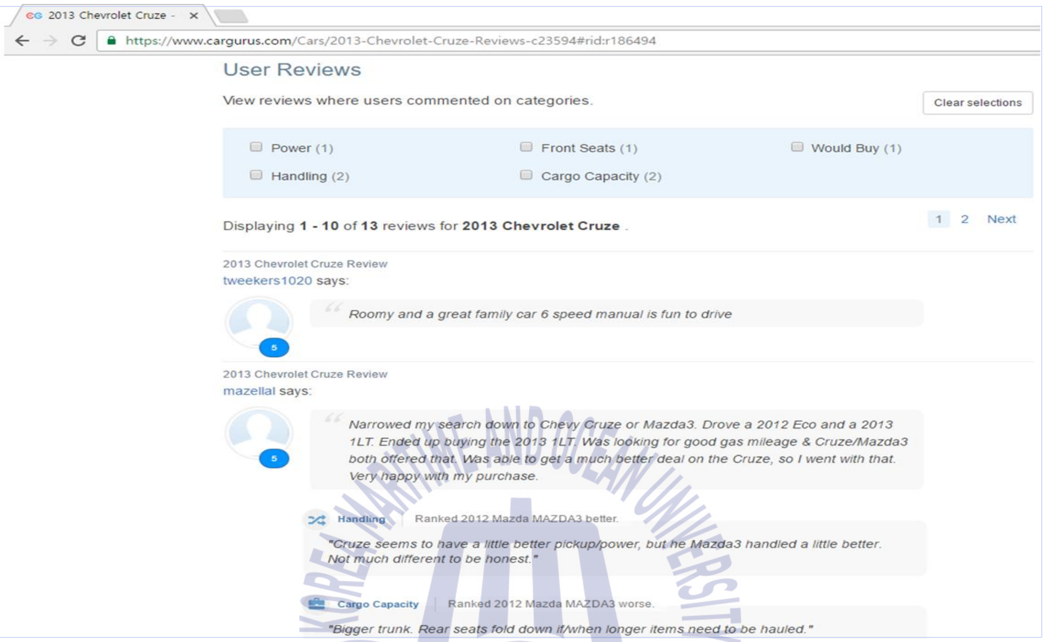
4.1.1 The Subject of Analysis & Crawling Data

In this study, the subjects for analysis are consumers' opinions, which are extracted the VOC of companies from "CarGurus" web site (www.cargurus.com) related to brand new and used vehicles in the U.S. They were consumers' opinions from 2013 to 2015. In this site, for the purpose of this present study, 45,447 pieces of customers' opinions on 26 car manufacturing companies had been collected including titles and status information. The opinions duplicated or written in other languages except for English were deleted from the sample data.

Table 3 Customers' Reviews

Company	2013	2014	2015	Total
GM	2303	2155	2203	6661
FCA	1624	1394	1666	4684
Ford	1145	1140	1484	3769
Honda	1912	1708	1835	5455
Volkswagen	964	690	656	2310
Others	7258	7546	7764	22568
Total	15206	14633	15608	45447

During the period, we could simply notice that the number of reviews reflect customers' interest in auto company with the number of reviews. GM obtained the most number of reviews as well as is the top of vehicle organizations in the U.S. Figure 9 is an example of user reviews and shows URL from the web site.



2013 Chevrolet Cruze -
https://www.cargurus.com/Cars/2013-Chevrolet-Cruze-Reviews-c23594#rid:r186494

User Reviews

View reviews where users commented on categories.

□ Power (1) □ Front Seats (1) □ Would Buy (1)

□ Handling (2) □ Cargo Capacity (2)

Displaying **1 - 10** of **13** reviews for **2013 Chevrolet Cruze** .

2013 Chevrolet Cruze Review

tweekers1020 says:

Roomy and a great family car 6 speed manual is fun to drive

2013 Chevrolet Cruze Review

mazellal says:

Narrowed my search down to Chevy Cruze or Mazda3. Drove a 2012 Eco and a 2013 1LT. Ended up buying the 2013 1LT. Was looking for good gas mileage & Cruze/Mazda3 both offered that. Was able to get a much better deal on the Cruze, so I went with that. Very happy with my purchase.

Handling | Ranked 2012 Mazda MAZDA3 better.

"Cruze seems to have a little better pickup/power, but he Mazda3 handled a little better. Not much different to be honest."

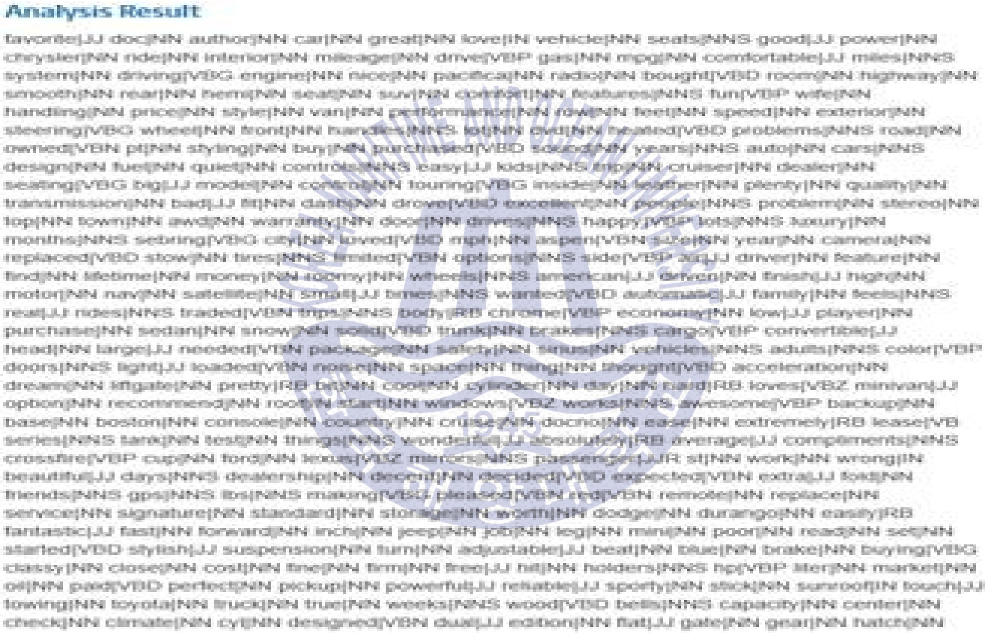
Cargo Capacity | Ranked 2012 Mazda MAZDA3 worse.

"Bigger trunk. Rear seats fold down if/when longer items need to be hauled."

Fig. 9 User Reviews

4.1.2 Extracting POS Information

It is possible that a word has a variety of meanings depending on POS (Part-of-Speech) in the text. For instance, as the word "good" when it is used as an adjective and when used as a noun, the polarity of the respective positive or negative meaning is also altered. Therefore, it is necessary to extract the POS information in order to analyze the positive or negative polarity with respect to customers' feedbacks. In this study, we took advantage of POS tagging at "www.TextAnalysisOnline.com" to extract the POS information of each word. TextAnalysis API (Application Programming Interface) provides customized Text Analysis or Text Mining Services like Word Tokenize, POS Tagging, Stemmer, Lemmatizer, Chunker, Parser, Key Phrase Extraction (Noun Phrase Extraction), Sentence Segmentation (Sentence Boundary Detection), Grammar Checker, Sentiment Analysis, Text Summarizer, Text Classifier and other Text Analysis Tasks. It stands on the giant shoulders of NLP Tools, such as NLTK, TextBlob, Pattern, MBSP and etc. TextAnalysisOnline displays the extracted the POS information as a format of XML. Figure 10 shows an example of extracting a POS information from words extracted by using Antconc 3.4.1w from the opinions posted by actual customers.



```
Analysis Result
favorite|JJ doc|NN author|NN car|NN great|NN love|IN vehicle|NN seats|NNS good|JJ power|NN
chrysler|NN ride|NN interior|NN mileage|NN drive|VBP gas|NN mpg|NN comfortable|JJ miles|NNS
system|NN driving|VBG engine|NN nice|NN pacifica|NN radio|NN bought|VBD room|NN highway|NN
smooth|NN rear|NN hemi|NN seat|NN suv|NN comfort|NN features|NNS fun|VBP wife|NN
handling|NN price|NN style|NN van|NN performance|NN row|NN feet|NN speed|NN exterior|NN
steering|VBG wheel|NN front|NN handles|NNS lot|NN dvd|NN heated|VBD problems|NNS road|NN
owned|VBN pt|NN styling|NN buy|NN purchased|VBD sound|NN years|NNS auto|NN cars|NNS
design|NN fuel|NN quiet|NN controls|NNS easy|JJ kids|NNS trip|NN cruiser|NN dealer|NN
seating|VBG big|JJ model|NN control|NN touring|VBG inside|NN leather|NN plenty|NN quality|NN
transmission|NN bad|JJ fit|NN dash|NN drove|VBD excellent|NN people|NNS problem|NN stereo|NN
top|NN town|NN awd|NN warranty|NN door|NN drives|NNS happy|VBP lots|NNS luxury|NN
months|NNS sebring|VBG city|NN loved|VBD mph|NN aspen|VBN size|NN year|NN camera|NN
replaced|VBD slow|NN tires|NNS limited|VBN options|NNS side|VBP 300|JJ driver|NN feature|NN
find|NN lifetime|NN money|NN roomy|NN wheels|NNS american|JJ driven|NN finish|JJ high|NN
motor|NN nav|NN satellite|NN small|JJ times|NNS wanted|VBD automatic|JJ family|NN feels|NNS
real|JJ rides|NNS traded|VBN trips|NNS body|RB chrome|VBP economy|NN low|JJ player|NN
purchase|NN sedan|NN snow|NN sold|VBD trunk|NN brakes|NNS cargo|VBP convertible|JJ
head|NN large|JJ needed|VBN package|NN safety|NN sirius|NN vehicles|NNS adults|NNS color|VBP
doors|NNS light|JJ loaded|VBN noise|NN space|NN thing|NN thought|VBD acceleration|NN
dream|NN liftgate|NN pretty|RB bit|NN cool|NN cylinder|NN day|NN hard|RB loves|VBZ minivan|JJ
option|NN recommend|NN roof|NN start|NN windows|VBZ works|NNS awesome|VBP backup|NN
base|NN boston|NN console|NN country|NN cruise|NN docno|NN ease|NN extremely|RB lease|VB
series|NNS tank|NN test|NN things|NNS wonderful|JJ absolutely|RB average|JJ compliments|NNS
crossfire|VBP cup|NN ford|NN lexus|VBZ mirrors|NNS passenger|JJR st|NN work|NN wrong|IN
beautiful|JJ days|NNS dealership|NN decent|NN decided|VBD expected|VBN extra|JJ fold|NN
friends|NNS gps|NNS lbs|NNS making|VBG pleased|VBN red|VBN remote|NN replace|NN
service|NN signature|NN standard|NN storage|NN worth|NN dodge|NN durango|NN easily|RB
fantastic|JJ fast|NN forward|NN inch|NN jeep|NN job|NN leg|NN mini|NN poor|NN read|NN set|NN
started|VBD stylish|JJ suspension|NN turn|NN adjustable|JJ beat|NN blue|NN brake|NN buying|VBG
classy|NN close|NN cost|NN fine|NN firm|NN free|JJ hit|NN holders|NNS hp|VBP liter|NN market|NN
oil|NN paid|VBD perfect|NN pickup|NN powerful|JJ reliable|JJ sporty|NN stick|NN sunroof|IN touch|JJ
towing|NN toyota|NN truck|NN true|NN weeks|NNS wood|VBD bells|NNS capacity|NN center|NN
check|NN climate|NN cyl|NN designed|VBN dual|JJ edition|NN flat|JJ gate|NN gear|NN hatch|NN
```

Fig. 10 An Example of POS Tagging

S (Sentence) / NN (Noun) / W (Word) / VG (Verb Group) C represents an attribute of the POS classification. There is a 30 kind of the part-of-speech classification such as PRP (Personal Pronoun), JJ (Adjective), NNS (Noun Plural), VBN (Verb, Past, Participle). The POS information identified in this way from each customer’s opinion was saved in the database. Through this POS tagging, we could classify POS of words, and extract noun and adjective vocabulary presented NN and JJ in the midst of these words arranged.

4.1.3 Review Extracting POS Information

We, at first, established lexicons of negative and positive emotions based on customers' feedbacks. By applying POS information extracted from VOC and extracting synonyms of words with WordNet, which offers the relation between words such as synonym and antonym, we expanded seed words. As the next step, we collected the polarity of the expended seed words from SentiWordnet and computed the sentimental intensity of each word. Thus, we could build the sentiment lexicon by selecting 4,815 negative and 2,021 positive sentiment words. Figure 11 shows how the process of the lexicon for negative and positive emotions was constructed. In this study, utilizing information related to synonyms, we expanded the seed words into lexicons as shown in table 4.

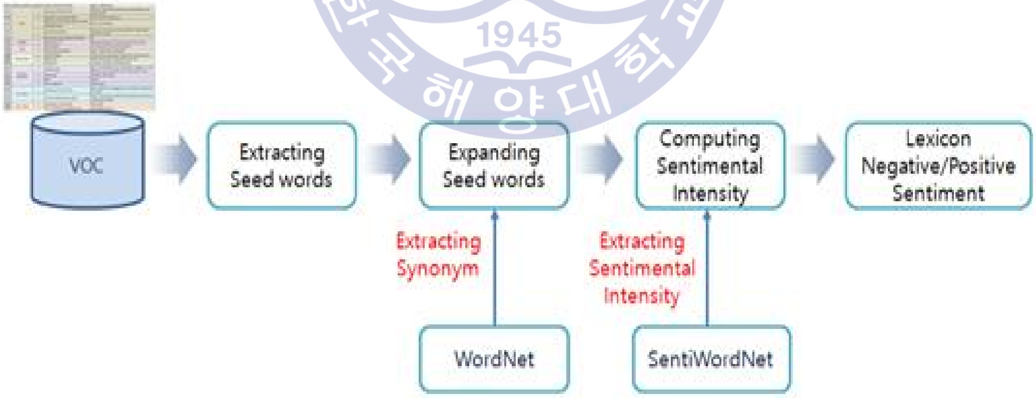

VOC
Extracting Seed words
Expanding Seed words
Computing Sentimental Intensity
Lexicon Negative/Positive Sentiment
Extracting Synonym
Extracting Sentimental Intensity
WordNet
SentiWordNet

Fig. 11 Process of Lexicon for Negative & Positive Emotions

Based on this sentiment lexicon, we expanded lexicon and classified negative and positive words into 7 sentiment domains; Sadness, Shame, Anger, Fear, Frustration, Delight, and Satisfaction. As we mentioned, the basic seed words including negative and positive emotions were extracted from the existing literature, but expanded the negative and positive sentiment words by applying car related words in table 4.

Table 4 Expanded Lexicon of Negative& Positive Domain

Domain	Seed words	Lexicon expanded
Sadness	sadness, loneliness, unhappiness, depression	sad, grief, sorrowful, mournful, lonely, lonesome, lone, unhappy, depressed, blue, slump and etc.
Shame	shame, guilt, regret, embarrassment	humiliation, humiliated, ashamed, mortified, dishonorable, shameful, regretful and etc.
Anger	anger, irritation, disgust, rage	anger, fury, resentment, wrath, indignation, angry, angered, fretful, annoyance, irritating, and etc.
Fear	fear, worry, anxiety, nervousness	dread, dreadful, dreaded, panic, fright, affright, alarm, aversion, fearfulness, fearful, fearsome and etc.
Frustration	frustration, resignation, powerlessness, despair	frustrated, frustrating, discouraged, disappointment, dejected, powerless, helpless and etc.
Delight	delight, pleasure, joy, happiness	pleased, pleasant, enjoyable, happy, ecstatic, enthusiastic, spirited, comfy, agile and etc.
Satisfaction	satisfaction, gratification, fulfillment	awesome, balanced, classy, sophisticated, stunning, sublime, refined, exquisite, impeccable and etc.

Based on these lexicons, even though they are categorized into the same range of emotion, the sentimental polarity of each word is different from one another. For instance, although "irritating," "angry" and "furious" are classified into the same domain of Anger, respective sentimental polarity differs from one another. This aspect means that words can have the intensity of a variety of emotions according to individual vocabulary, usage, context, POS, or even in the same POS. In short, a word can have multiple meanings. SentiWordNet 3.0 offers the polarity information indicating the emotional intensity for each word depending on the POS and usage, and also provides an index information indicating the frequency of use (Diener et al., 1995). The index represents the rank in accordance with the frequency of use. In this study, we calculated the sentiment intensity more accurately and precisely taking into account the index information. For example, in table 5, the word "sad" used as an adjective is shown from SentiWordNet 3.0 and figure 12 shows classifying polarity of the word "sad" into three aspects based on the frequency of use.

Table 5 Polarity Information of Word "Sad"

Word	POS	Index Value	Positivity	Negativity
sad	Adj.	1	0.125	0.75
		2	0	0.25
		3	0	1

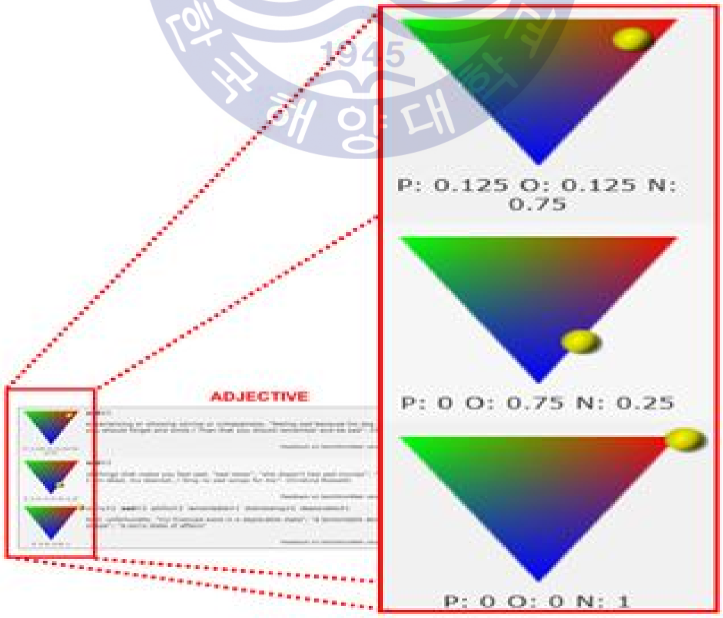


Fig. 12 Polarity Information from SentiWordNet 3.0

To begin with, we respectively calculated difference between positivity and negativity of each meaning, and then multiplied the reciprocal of the index values given to its meaning. The higher the index value becomes, the greater weight is assigned. Once for all, we normalized their total sums. That is, in the case of "sad", the negativity value was extracted using this example of formula (Jung, 2013).

$$\{(0.75-0.125)*1+(-0.25)*(1/2)+(-1)*(1/3)\}/\{(1 + (1/2) + (1/3)\} \qquad (1)$$

Through text matching both the extracted negativity and positivity values and lexicon constructed by the process above, intensity of sentiment of each VOC are measured. Based on these procedure, we constructed the sentiment lexicon selecting 4,815 negative and 2,021 positive sentiment words and built the 7 sentiment domains. It is meaningful to construct the sentiment lexicon related to automative industry.

4.2 Phase II : Reliability Analysis

Phase II is a research step to determine whether the sentiment of the auto sales market includes predictable information and validate whether 7 sentiment domains are appropriately classified into negativity and positivity.

Based on the sentiment lexicon, we combined these extracted and classified negative and positive words with words related to the automobile industry, and analyzed a total of 45,474 pieces of customers' opinions of 26 car manufacturing companies had been crawled from Jan. 1. 2013 to Dec. 31. 2015. In the middle of 45,457 threads in data set from 26 brands, the threads were discussed 229 unique vehicle models about their products and services. The average thread contained 4 sentences with a total 16 words (min 1 word; max 272 words). In the table 6, it demonstrates basic descriptive statistics on the amounts of 1-month sentiment change during 3 years. The amounts of 1-month sentiment change(E_t) is measured according to this following formula (O'Leary, 2011).

$$E_t = \ln(S_t / S_{t-1}) \qquad (2)$$

"ln" is natural logarithm. S_t is the frequency of this month sentiments and $S_{t-1}$is the frequency of the previous month's sentiments. The reason why we use the amount of sentiment change as variable is that, basically, the distribution of sentiment shows a similar aspect (Jung & Nah, 2007). The similar distribution of emotion responds the varied change such as sales or market share. In addition, since the number of posting shows a large difference depending on product or brand, the rate of emotion, that is, the amount of sentiment change is appropriate for variable.

The average of Sadness, Shame, Anger, Fear, and Frustration, indicating the negativity of the market, is having a positive value, while that of Delight and Satisfaction, presenting the positivity of it, has a negative value. We can ascertain that the domains of sentiment properly divided as the average of emotion of other propensity simultaneously have the opposite sign to each other (Table 6).

Table 6 Basic Descriptive Statistics

Variables	Min.	Max	Aver.	St. Dev
Sadness	-0.7826	0.4521	0.000401	0.1115076
Shame	-1.8064	1.1331	0.001240	0.1632425
Anger	-0.7207	0.8998	0.001341	0.1423444
Fear	-0.9557	1.1148	0.000892	0.1457385
Frustration	-0.9897	0.6557	0.000902	0.1471391
Delight	-0.9673	0.6099	-0.000531	0.1425756
Satisfaction	-0.8895	1.3387	-0.000684	0.1916136

4.2.1 Auto-correlation Analysis of Sentiment

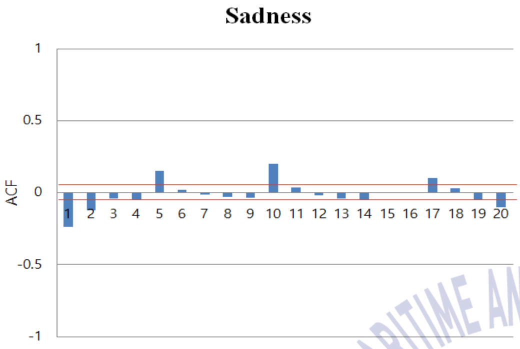
In this study, we conducted the auto-correlation analysis to determine whether the sentiment of the auto sales market includes predictable information. As dealing with time-series data, It is always likely that the continuous error terms are correlated to each other. In some particular time point, the error term at that time includes not only the impact of that time, but the influence transferred from the impact from the past. Because of these transferred influence, the impact at that time are correlated with the impact from the past and this situation produces correlation the error terms. In this case, auto-correlation exists and auto-correlation of both positives and negative could exist. At first, we empirically analyzed whether the change of each emotion has the features of auto-correlation. If the auto-correlation is identified, it suggests that sentiment containing information about the auto sales market may actually have an influence on the market. Table 7 is presenting the result of analyzing the auto-correlation of 20 months in the midst of 36 months, the 20 months were randomly selected arranged in order of month as lags and it is a summary on the auto-correlation coefficient of 7 sentiment domains shown in the car community. AC denotes an auto-correlation coefficient, and Q-Stat. means Ljung-box statistics. Value is the value of Ljung-box statistics can confirm accompanying P-value (Sig.). We omitted the p-value (<0.01) under Q-stat in the table 7 as it is respectively significant in the each sentiment domain (0.0000*). The Ljung-Box test (named for Greta M. Ljung and George E. P. Box) known as the Ljung–Box Q test is a type of statistical test of whether any of a group of auto-correlations of a time series are different from zero. Instead of testing randomness at each distinct lag, it tests the "overall" randomness based on a number of lags (Ljung & Box, 1978).

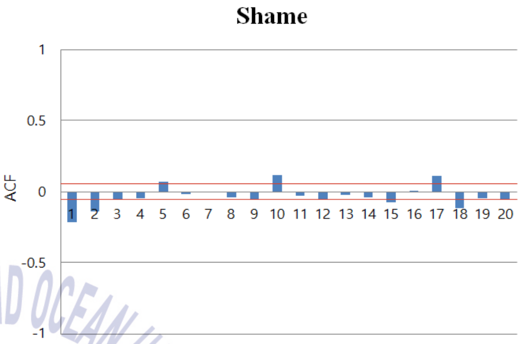
Table 7 The Result from Auto-correlation Analysis

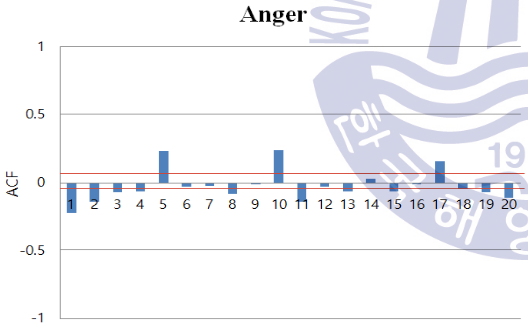
	Sadness		Shame		Anger		Fear		Frustration		Delight		Satisfaction	
Lag	AC	Q-stat. value	AC	Q-stat. value	AC	Q-stat. value	AC	Q-stat. value	AC	Q-stat. value	AC	Q-stat. value	AC	Q-stat. value
1	-0.238	41.748	-0.212	33.226	-0.221	35.665	-0.123	11.765	-0.224	37.103	-0.182	24.174	-0.175	22.644
2	-0.127	53.697	-0.137	43.564	-0.143	50.925	-0.212	33.665	-0.147	52.769	-0.179	43.572	-0.161	41.456
3	-0.044	55.155	-0.054	49.433	-0.072	53.697	-0.084	39.736	-0.064	55.155	-0.039	47.515	-0.099	48.155
4	-0.046	56.732	-0.045	52.497	-0.066	57.732	-0.033	46.701	-0.077	56.732	0.004	61.732	-0.006	63.343
5	0.148	67.792	0.072	62.754	0.235	87.721	0.168	64.219	0.186	77.761	0.242	96.779	0.268	102.424
6	0.018	67.991	-0.013	67.098	-0.031	89.959	0.022	68.442	0.008	79.265	-0.081	107.271	0.023	104.931
7	-0.016	74.209	-0.006	72.272	-0.025	94.214	0.002	72.529	-0.046	81.233	-0.048	112.742	-0.021	124.331
8	-0.034	75.052	-0.039	75.452	-0.084	95.873	-0.054	78.108	-0.023	85.052	-0.024	115.156	-0.029	125.398
9	-0.038	75.113	-0.058	76.738	-0.011	95.298	-0.048	79.073	-0.061	87.511	-0.045	117.411	-0.041	125.882
10	0.197	77.128	0.121	77.309	0.239	106.185	0.147	84.643	0.176	92.128	0.209	120.612	0.201	132.318
11	-0.005	79.147	-0.029	77.647	-0.142	108.429	0.001	88.292	-0.005	96.153	-0.012	123.761	-0.027	136.539
12	-0.068	79.566	-0.055	82.069	-0.032	109.786	-0.062	92.433	-0.048	99.426	-0.084	133.795	-0.071	139.245
13	-0.044	79.727	-0.022	83.732	-0.067	110.732	-0.041	99.749	-0.091	102.325	-0.047	137.977	-0.049	141.778
14	-0.046	79.929	-0.036	89.425	0.026	117.492	-0.006	102.11	-0.024	109.467	-0.001	139.232	-0.051	142.929
15	-0.005	82.132	-0.075	96.774	-0.065	122.479	-0.017	114.54	-0.075	122.352	-0.062	149.509	-0.013	172.236
16	-0.002	82.821	0.003	97.672	-0.012	123.217	0.009	122.8	0.011	125.284	0.002	152.828	-0.102	178.821
17	0.101	83.242	0.112	103.002	0.156	125.739	0.052	124.4	0.033	128.257	0.090	158.322	0.081	182.134
18	-0.029	83.544	-0.113	104.396	-0.042	133.692	-0.094	124.61	-0.039	128.954	-0.032	161.201	-0.089	183.418
19	-0.056	84.272	-0.042	107.465	-0.071	147.431	-0.049	128.43	-0.075	133.842	-0.055	164.423	-0.097	189.412
20	-0.102	88.798	-0.054	108.175	-0.111	168.524	-0.032	128.99	-0.009	138.751	-0.072	168.336	-0.102	192.067

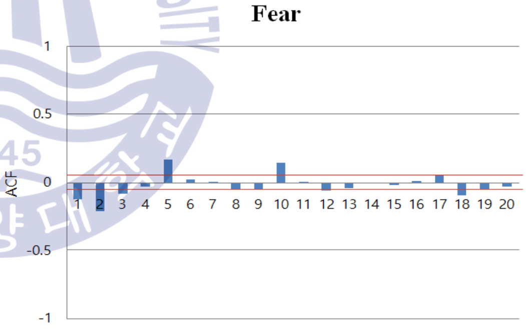
In table 7, we can grasp from this results thar the sentiment data of VOC have a serial pattern. Although there are slight differences in respective sentiment domain, the present sentiment has a significant amount of AC (auto-correlation coefficient) on a cycle of 5 and 7 months. According to "Consumer Report" and "U.S. News & World Report" (from 2013 to 2015), the time of new cars released is mostly fixed in May and October because it depends on the buying season. In May, auto companies focus on the summer vacation season and in the early October, they aim to selling of the end of the year. In particular, manufacturers often then begin production on the next year's models during the late fall and early winter months - often a slower time for vehicle sales. Once the newer models are released, most dealerships host both the new and older models on sales lots and, in many cases, feature the older models at a discounted price during the busy shopping months of October, November and December. At the time, customers evaluate the vehicles which they purchased and then post their reviews on the web such as car e-communities or blogs. Thus, we infer that it is likely that their emotions of vehicle are perceived and the amounts of sentiment change could be increased during this period of time.

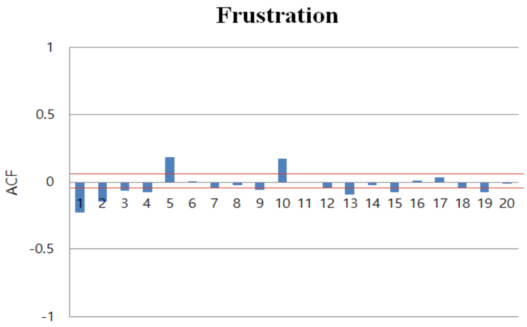
We also can recognize that the time difference exceeding the confidence limit(-0.05<ACF<0.05) exists in sentiment domain. Figure 13 demonstrates ACF (auto-correlation function) of 7 domains from 1 month to 20. As X-axis is Lag number and Y-axis means auto-correlation coefficients, the reference lines up and down based on 0 criteria display confidence limits. Similarly to table 7, through the figure 13, it can be seen that the time difference exists out of the confidence limit. That is, the present 7 sentiment domains are used to predict the automotive market share of the future and we can identify that the relation between the past value and the current is significant. This aspect shows that the current sensibility can be used to predict the future and the relationship between the past values and the current is significant. In other words, it implies that 7 emotions on the automotive market are not meaningless values to be produced at random, but have features of information with the predictability and the period.

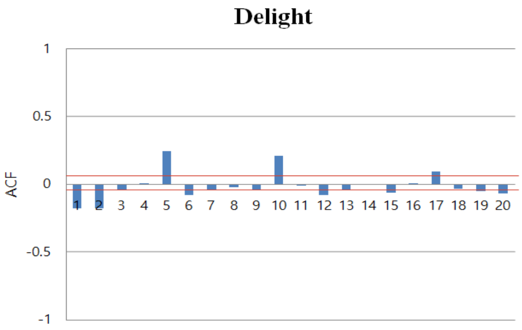

Sadness


Shame


Anger


Fear


Frustration


Delight

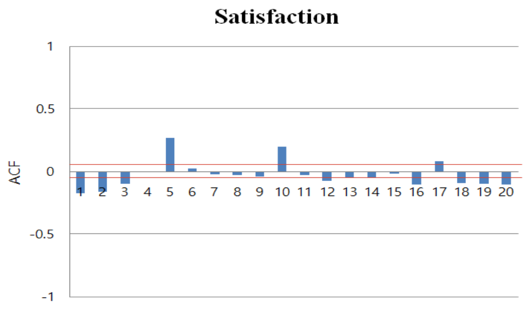


Fig. 13 Auto-correlation Analysis

In this study, we validated and verified the fact that the time-series information exists in the VOC sentiment data from the result through auto-correlation analysis. However, we didn't conduct a forecasting experiment under the judgment that it is impossible to predict the sales market share of the automobile market with only sentiment information in terms of prediction.

4.2.2 Principal Component Analysis of Sentiment

The purpose of this analysis is to extract some main factors to be able to account for most of the variance of the total of samples. The principal component extraction can be based on the value of Eigen-value 1 or more, which is the sum of squares of loadings. However, in this study, we fixed the number of components into 3 in advance in order to evaluate the sentiment on market by separating the three aspects; positivity, negativity and neutrality. As the component rotation was conducted with Varimax 20times repeatedly to extract principal component vector, which is most commonly used in the orthogonal component rotation system and maximizes the sum of the variances of the squared loadings (squared correlations between variables and components).

The rotated component matrix of 7 sentiment domains is shown in table 8. Sadness, Anger, and Fear has been divided into a principal component 1. Shame and Frustration has been divided into the principal component 2. Delight and Satisfaction is connected to the principal component 3. The principal component 1 represents the negativity on the auto market such as Sadness, Anger, and Fear. The principal component 2 includes neutrality. The principal component 3 including Delight and Satisfaction mainly shows the positivity on the market. The market asymmetrically responds depending on customer's feedback, whether it is negative or positive.

Table 8 Rotated Component Matrix

Domain	PC1	PC2	PC3
Sadness	0.672	0.294	0.223
Shame	0.197	0.742	0.049
Anger	0.781	0.382	0.012
Fear	0.645	0.227	0.307
Frustration	0.201	0.798	0.154
Delight	0.077	0.559	0.726
Satisfaction	0.059	0.329	0.851

The negative and positive data on vehicles classified in this study will be able to have a different impact on the market performance. In other words, It suggests that the emotion plays a role in the information to predict the change of the car sales in the auto market through the negativity and positivity on cars. In particular, Shame and Frustration are classified with the negative emotion in our own lexicon. However, as a result of the PCA, Shame and Frustration have not been included in any group of the negative or the positive. Thus, it contains the neutral information or is regarded as the noise regardless of the plus (+) or minus (-) on the auto sales. In case of Delight, although it includes the neutrality and the positivity, as the main component is the positivity, it is included in the positive domain. Technically, it is difficult to extract exactly the information we want in the internet space with a lot of information through a combination of words. We reckon that it is necessary to analyze additionally the sentiment on the respective domain of Shame and Frustration mechanically classified.

4.3 Phase III : Influence on Automotive Market Share

4.3.1 Linear Regression Model

In the research model, we examined how VOC, which is consumer review on the products of the car manufacturing companies to target the online community, one of the social media channels, has an impact on the automotive market share. VOC on the products of automobile manufacturing companies is posted in a format of text. The independent variables of this study was converted to the quantified sentiment domains of texts by applying the opinion mining techniques. In the community for these products, we graphically represent two models to explain the ups and downs of the market share on VOC digitizing sensibility by applying the opinion mining as follows.

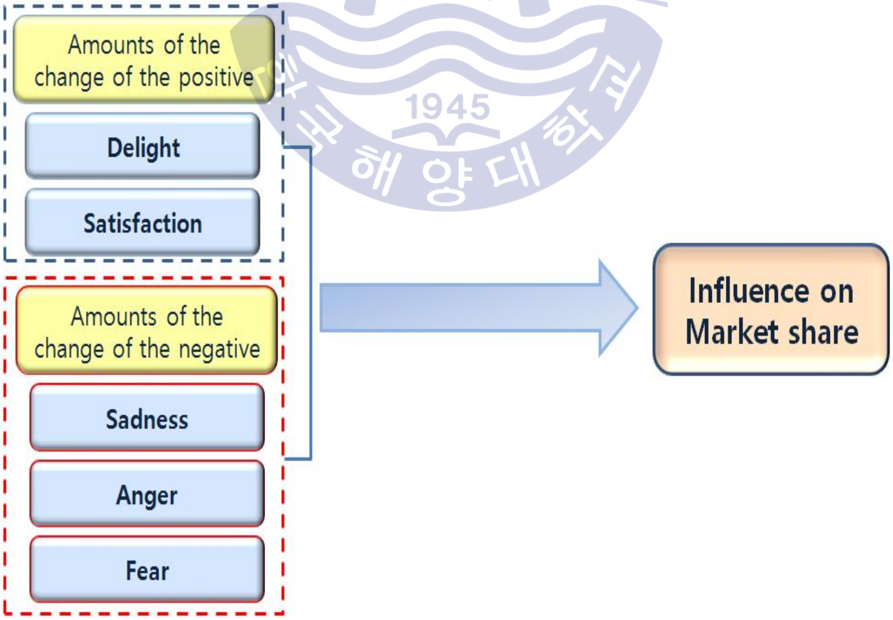

Amounts of the change of the positive
Delight
Satisfaction
Amounts of the change of the negative
Sadness
Anger
Fear
Influence on Market share

Fig. 14 Linear Regression Analysis Model 1

In linear regression analysis model 1 (Figure 14), we fixed five variables, the amounts of change the positive and the negative, as independent variables. According to PCA, the sentiment domains inclining to the principal component of positivity are "Delight" and "Satisfaction" and those of negativity are "Sadness," "Anger," and "Fear". The amounts of the change of these two positive domains and three negative domains become independent variables.

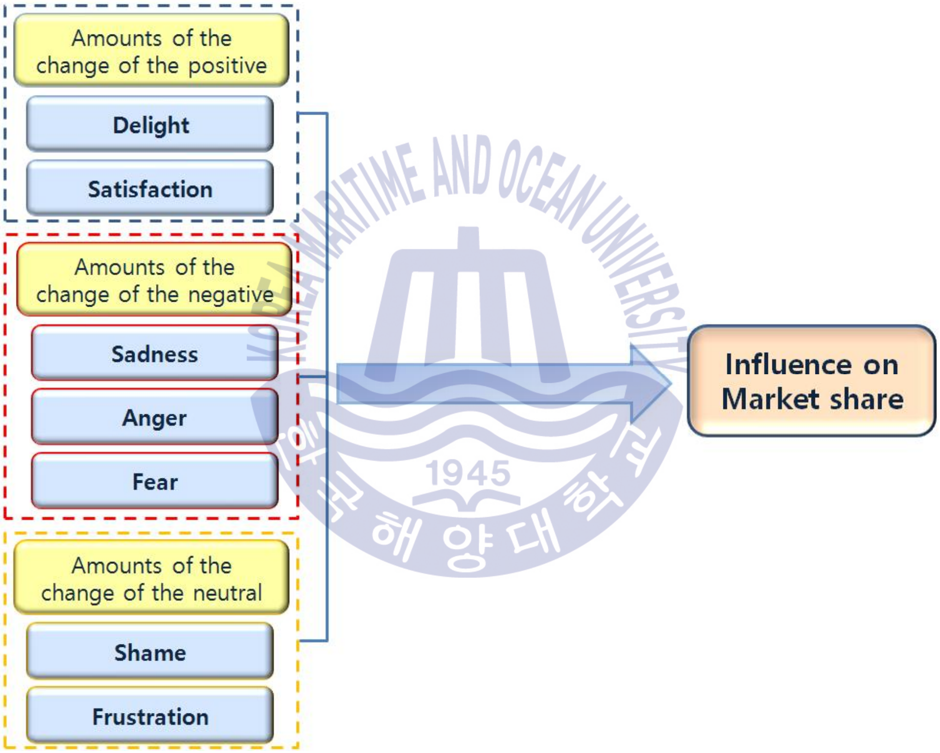

Amounts of the change of the positive
Delight
Satisfaction
Amounts of the change of the negative
Sadness
Anger
Fear
Amounts of the change of the neutral
Shame
Frustration
Influence on Market share

Fig. 15 Linear Regression Analysis Model 2

In Model 2 (Figure 15), we added the amounts of change of the neutral as independent variables in order to investigate what the neutrality of sentiment effects on the market share of the company. As a result of PCA, two domains, "Shame" and "Frustration", are independent variables including the neutrality. In addition, even though the domain "Delight" belongs to the positive domain, it simultaneously is inclined to both neutrality and positivity. It is necessary to focus on the result of analysis of "Delight," and "Shame" and "Frustration."

4.3.2 Definition of Variables

Independent variables is the amounts of 1-month sentiment change based on the frequency of each sentiment domain. As mentioned in chapter 3, the amounts of 1-month sentiment change (E_t) is measured according to this following formula (O'Leary, 2011).

$$E_t = \ln(S_t / S_{t-1}) \tag{2}$$

ln: natural logarithm.

S_t: The frequency of this month sentiments

S_{t-1}: The frequency of the previous month's sentiments.

After we classified 7 sentiment domains and selected four major organizations such as GM Group, FORD, FCA and VOLKSWAGEN, we measured the amounts of sentiment change of 7 sentiment domains (Sadness, Shame, Anger, Fear, Frustration, Delight and Satisfaction) for each company as independent variable.

Dependent variable was selected for automative market share. As shown in table 9, we investigated the monthly vehicle market share from 2013 to 2015 through Automakers & ANDC, and 4 companies collected from 26 car manufacturing companies, which are presenting the meaningful change in market share.

Table 9 Monthly Vehicle Market Share in the U.S.

Month	GM	FORD	FCA	VOLKSWAGEN
Jan-2013	0.187	0.159	0.113	0.041
Feb-2013	0.188	0.164	0.117	0.038
Mar-2013	0.169	0.162	0.118	0.037
Apr-2013	0.185	0.165	0.122	0.040
May-2013	0.175	0.170	0.115	0.038
Jun-2013	0.189	0.167	0.112	0.039
Jul-2013	0.178	0.147	0.107	0.040
Aug-2013	0.184	0.147	0.110	0.039
Sep-2013	0.165	0.162	0.126	0.042
Oct-2013	0.188	0.159	0.116	0.037
Nov-2013	0.170	0.153	0.114	0.039
Dec-2013	0.169	0.159	0.118	0.040
Jan-2014	0.170	0.152	0.126	0.036
Feb-2014	0.186	0.154	0.130	0.035
Mar-2014	0.167	0.158	0.126	0.036
Apr-2014	0.183	0.151	0.128	0.036
May-2014	0.177	0.157	0.121	0.033
Jun-2014	0.188	0.156	0.120	0.035
Jul-2014	0.178	0.147	0.117	0.035
Aug-2014	0.172	0.152	0.125	0.036
Sep-2014	0.179	0.144	0.136	0.036
Oct-2014	0.177	0.147	0.133	0.038
Nov-2014	0.173	0.143	0.131	0.041
Dec-2014	0.182	0.145	0.128	0.038
Jan-2015	0.176	0.154	0.126	0.034
Feb-2015	0.184	0.143	0.130	0.032
Mar-2015	0.162	0.152	0.128	0.037
Apr-2015	0.185	0.152	0.130	0.036
May-2015	0.179	0.153	0.124	0.035
Jun-2015	0.176	0.152	0.125	0.036
Jul-2015	0.180	0.147	0.118	0.036
Aug-2015	0.171	0.148	0.128	0.036
Sep-2015	0.174	0.153	0.134	0.033
Oct-2015	0.181	0.146	0.134	0.035
Nov-2015	0.173	0.143	0.131	0.041
Dec-2015	0.177	0.145	0.132	0.032

Source: Automakers & ANDC

4.3.3 The Result of Linear Regression Analysis

1) GM group

In GM's case, it had been the top auto manufacturing organization during this period (2013-2015) according to Automakers & ANDC. Although, on March, 2013 (16.9%), September, 2013 (16.5%), March, 2014 (16.7%) and March, 2015 (16.2%), the radical fluctuations of market share were shown, it had maintained the top of market share in the US. The results of regression investigating the impact on the market share in accordance with the change of sentiment in VOC are stated as follows.

R^2 means statistics, which describes how much the independent variables explain the dependent variable. In table 10 since R^2 is 0.363, 5 sentiment domains can explain the market share approximately 36.3%. As adjusted R^2 is 0.357, it shows similar level to R^2. As Durbin-Watson (DW) value shows from 1 to 3, it is unproblematic to independency of residual. Since the DW value is 1.56, we reason that the condition of the independency of residual is met. In ANOVA, investigating P-value (significant probability) on F-value, as P-value is 0.015, it is appropriate for the regression model 1.

As the result of coefficients, generally, if VIF is less than 10, it is unproblematic to multi-collinearity. VIF represents the level between 1 and 3 in the table 10.

Table 10 The Result of Model 1 (GM Group)

Model	R^2	Adjusted R^2	Durbin-Watson	ANOVA		Variables	Unstrdzd. Coefficient		Strdzd. Coeffi.	t	sig.	Collinearity Stat.
				F	P-v.		B	Std. Error	B			VIF
1	0.363	0.357	1.560	3.416	0.015	(constant)	0.215	0.017		12.819	0.000	
						Sadness	-0.211	0.087	**-0.492**	-2.425	**0.022**	1.937
						Anger	-0.004	0.010	**-0.466**	-2.430	**0.067**	1.120
						Fear	-0.078	0.132	-0.122	-0.593	0.258	1.985
						Delight	0.181	0.324	**0.159**	2.557	**0.058**	1.820
						Satisfaction	0.014	0.241	0.017	-0.059	0.353	2.667

As the result of significancy of each variable, in domain Sadness, Anger, and Delight, P-value is shown to respectively 0.022, 0.067, and 0.058. It represents that Sadness, Anger, and Delight of 5 domains to be measured have an significant influence on the GM's performance. In addition, as we see the standardized coefficient, Sadness (-0.492) and Anger (-0.466) have negative (-) influence on market share and Delight has a positive (+) impact on the market share of GM, but it is a little meager level (0.159).

In table 11 since R^2 is 0.403, 7 sentiment domains can explain the market share approximately 40.3%. As adjusted R^2 is 0.354, it shows similar level to R^2. Since the DW value is 1.64, we can evaluate that the condition of the independency of residual is met. In ANOVA, investigating P-value on F-value, as P-value is 0.029, it is suitable for the regression model 2. As the result of coefficients, generally, it is unproblematic to multi-collinearity as VIF represents below 10 in the table 11.

Table 11 The Result of Model 2 (GM Group)

Model	R^2	Adjusted R^2	Durbin-Watson	ANOVA		Variables	Unstrdzed. Coefficient		Strdzd. Coeffi.	t	sig.	Collinearity Stat.
				F	P		B	Std. Error	B			VIF
2	0.403	0.354	1.640	2.701	0.029	(Constant)	0.222	0.021		10.433	0.000	
						Sadness	-0.238	0.111	**-0.555**	-2.147	**0.041**	1.634
						Anger	0.002	0.011	0.028	0.180	0.259	1.332
						Fear	-0.089	0.136	-0.139	-0.658	0.112	1.922
						Delight	-0.212	0.326	-0.187	-0.652	0.120	2.731
						Satisfaction	0.070	0.249	0.061	0.279	0.282	2.667
						Shame	0.142	0.189	0.183	0.749	0.106	2.723
						Frustration	-0.233	0.190	-0.204	-1.222	0.232	2.112

As the result of significancy of each variable, P-value of domain Sadness (0.041) has only an significant influence on the GM's performance. In addition, as we see the standardized coefficient, Sadness (-0.555) has an negative (-) influence on market share of GM. Domains Shame and Frustration, evaluated possessing neutrality by PCA, are not significant.

In case of GM, we could reason that Sadness of sentiment domains on customer reviews had negatively impacted on the business performance on a significantly high level (-0.492 in model 1 and -0.555 in model 2), compared to the market share from January, 2013 to December, 2015 (from max 18.9% to min 17.7%).

2) FORD

In FORD's case, it had retained the second auto manufacturing organization during this period('2013-'2015) according to Automakers &

ANDC. There were sharp declines on July. 2013 (0.147), July. 2014 (0.147) and at the early of 2015, the radical fluctuations of market share were shown such as January. 2015 (0.154), February. 2015 (0.143), and March. 2015(0.152). It, however, had maintained the second of market share in the US. The results of regression investigating the impact on the market share in accordance with the change of sentiment in VOC are stated as follows.

In table 12, as R^2 is 0.234, 5 sentiment domains can explain the market share approximately 23.4%. DW value (1.935) meets the condition of the independency of residual. In ANOVA, P-value (0.013) measured on F-value is suitable for the regression model 1. In parameter estimates, As the result of coefficients, generally, it is unproblematic to multi-collinearity as VIF represents below 10 in the table 12.

Table 12 The Result of Model 1 (FORD)

Model	R^2	Adjusted R^2	Durbin-Watson	ANOVA		Variables	Unstrdzed. Coefficient		Strdzd. Coeffi.	t	sig.	Collinearity Stat.
				F	P-v.		B	Std. Error	B			VIF
1	0.234	0.206	1.935	1.831	0.013	(constant)	0.149	0.019		7.872	0.000	
						Sadness	-0.012	0.009	**-0.218**	-1.317	**0.058**	1.073
						Anger	-0.265	0.151	**-0.332**	-1.756	**0.059**	1.397
						Fear	0.201	0.211	0.215	0.953	0.348	1.987
						Delight	0.367	0.183	**0.467**	2.005	**0.054**	2.111
						Satisfaction	-0.230	0.156	-0.655	-1.474	0.151	7.727

As the result of significancy of each variable, P-value of domains Sadness (0.058), Anger (0.059) and Delight (0.054) have significant influence on the FORD's market share. In addition, as we see the standardized coefficient, Sadness (-0.218) and Anger (-0.332) have negative (-) influence on market share of FORD while Delight (0.467) affects on it on a bit high level.

In table 13, we can realize that 7 sentiment domains can explain the market share approximately 23.7% through R^2. Since the DW value is 1.925, we are able to evaluate that the condition of the independency of residual is met. In ANOVA, investigating P-value (0.023) on F-value is suitable for the regression model 2. As the result of parameter estimates, VIF represents that multi-collinearity is stable between 1 and 3 except Satisfaction (12.047). Domain Satisfaction has a problem to multi-collinearity. It occurs since a similar variable is included in the independent variable.

Table 13 The Result of Model 2 (FORD)

Model	R^2	Adjusted R^2	Durbin-Watson	ANOVA		Variables	Unstrdzed. Coefficient		Strdzd. Coeffi.	t	sig.	Collinearity Stat.
				F	P		B	Std. Error	B			VIF
2	0.237	0.227	1.925	1.244	0.023	(Constant)	0.145	0.030		4.812	0.000	
						Sadness	-0.012	0.010	**-0.224**	-1.276	**0.065**	1.127
						Anger	-0.269	0.157	**-0.336**	-1.717	**0.067**	1.405
						Fear	0.202	0.222	0.216	0.908	0.372	2.068
						Delight	0.391	0.203	0.431	1.922	0.213	1.553
						Satisfaction	-0.263	0.201	-0.748	-1.306	0.202	12.047
						Shame	-0.040	0.159	-0.043	-0.251	0.204	1.077
						Frustration	0.088	0.335	0.065	0.262	0.195	2.269

As the result of significancy of each variable, P-value of domains Sadness (0.065) and Anger (0.067) has slightly an significant influence on the FORD's market share. It shows the similar aspect to model 1. In addition, as we see the standardized coefficient, Sadness (-0.224) and Anger (-0.336) have negative (-) influence on the market share of FORD. Compared to between the development of market share and the sentiment from 2013 to 2015, the market share had reduced 2.7% point from 17% to 14.3%.

3) FCA

FCA's market share had inclined approximately 2% point from 11.3% to 13.2% during this period (2013-2015) according to Automakers & ANDC. In the middle of 2013, it had dropped to 10.7% (July). After that, however, it steadily had upturned and retained the range of 13% at the end of 2015. The results of regression investigating the impact on the market share in accordance with the change of sentiment in VOC are stated as follows.

Table 14 The Result of Model 1 (FCA)

Model	R^2	Adjusted R^2	Durbin-Watson	ANOVA		Variables	Unstrdzed. Coefficient		Strdzd. Coeffi.	t	sig.	Collinearity Stat.
				F	P-v.		B	Std. Error	B			VIF
1	0.279	0.259	1.911	2.322	0.038	(constant)	0.124	0.022		5.675	0.000	
						Sadness	-0.389	0.153	**-0.255**	-2.536	**0.017**	1.993
						Anger	0.046	0.183	0.059	0.254	0.802	2.230
						Fear	0.229	0.153	0.301	1.492	0.146	1.688
						Delight	0.105	0.238	**0.354**	2.442	**0.061**	2.041
						Satisfaction	0.022	0.268	**0.328**	3.483	**0.034**	1.646

In table 14, as R^2 is 0.279, 5 sentiment domains can explain the market share approximately 27.9%. DW value (1.911) meets the condition of the independency of residual. In ANOVA, P-value (0.038) measured on F-value is suitable for the regression model 1. In parameter estimates, VIF is unproblematic to multi-collinearity.

As the result of significancy of each variable, P-value of domains Sadness (0.017), Delight (0.061) and Satisfaction (0.034) have significant influence on the FCA's market share. As seeing the standardized coefficient, Sadness (-0.225) has a negative (-) influence on market share of FCA (0.154) and Delight (0.354) and Satisfaction (0.328) do positively (+).

According to analysis on Model 2 (FCA), 7 sentiment domains can explain the market share approximately 35.1% through R^2. In ANOVA, P-value(0.03) measured on F-value is appropriate for the regression model 2. In parameter estimates, considering VIF, it is unproblematic to multi-collinearity.

Table 15 The Result of Model 2 (FCA)

Model	R^2	Adjusted R^2	Durbin-Watson	ANOVA		Variables	Unstrdzed. Coefficient		Strdzd. Coeffi.	t	sig.	Collinearity Stat.
				F	P		B	Std. Error	B			VIF
2	0.351	0.228	1.159	2.159	0.030	(Constant)	0.140	0.023		5.985	0.000	
						Sadness	-0.363	0.152	**-0.417**	-2.392	**0.024**	2.015
						Anger	0.036	0.180	0.046	0.202	0.342	2.233
						Fear	0.387	0.176	**-0.508**	2.197	**0.036**	2.304
						Delight	0.112	0.242	**0.364**	2.462	**0.038**	5.408
						Satisfaction	0.011	0.274	0.013	0.939	0.169	5.022
						Shame	-0.030	0.196	-0.033	-1.155	0.178	1.897
						Frustration	-0.361	0.274	-0.320	-1.318	0.198	2.551

As the result of significancy of each variable, it assumes different aspect from the result of model 1. P-value of domains Sadness (0.024), Fear (0.036) and Delight (0.038) have significant influence on the FCA's market share. We are able to infer that some positive and negative factors of Shame and Frustration impact on other sentiment domains. As seeing the standardized coefficient, Sadness (-0.417) and Fear (-0.508) have negative (-) influence on market share of FCA and Delight (0.364) does positively (+).

Compared to between the development of market share and the sentiment from 2013 to 2015, the market share had increased 2.1% point from 11.3% to 13.4%. In case of FCA, even though negativity and positivity had coexisted during the period, we are able to infer that the positive sentiment factors such as Delight and Satisfaction in VOC had impacted on the business performance of FCA.

4) VOLKSWAGEN

In VOLKSWAGEN's case, it had retained the range of market share between about 3% and 4% during this period (2013-2015) according to Automakers & ANDC. Except for the radical decrease (-0.009) of market share in December. 2019 (0.032), it shows steady rise and fall without dramatic fluctuation in spite of "Diesel gate," which is the event of exhaust fabrication. The results of regression investigating the impact on the market share in accordance with the change of sentiment in VOC are stated as follows.

In table 16, as R^2 is 0.33, 5 sentiment domains can explain the market share approximately 33%. DW value (1.541) meets the condition of the independency of residual. In ANOVA, P-value (0.043) measured on F-value is suitable for the regression model 1. In parameter estimates, VIF, below 10, is unproblematic to multi-collinearity.

Table 16 The Result of Model 1 (VOLKSWAGEN)

Model	R^2	Adjusted R^2	Durbin-Watson	ANOVA		Variables	Unstrdzed. Coefficient		Strdzd. Coeffi.	t	sig.	Collinearity Stat.
				F	P-v.		B	Std. Error	B			VIF
1	0.330	0.309	1.541	3.735	0.043	(constant)	0.051	0.013		3.975	0.000	
						Sadness	-0.003	0.035	-0.018	-0.099	0.222	1.065
						Anger	-0.110	0.200	**-0.343**	-2.551	**0.059**	2.250
						Fear	-0.109	0.061	-0.326	-1.797	0.082	1.107
						Delight	0.018	0.092	0.053	0.191	0.250	2.557
						Satisfaction	0.031	0.072	0.110	0.432	0.269	2.189

As the result of significancy of each variable, P-value of domain Anger (0.059) has a significant influence on the performance of VOLKSWAGEN. As observing the standardized coefficient, Anger (-0.343) has a negative (-) influence on market share of VOLKSWAGEN.

In table 17, we can realize that 7 sentiment domains can explain the market share approximately 55.1% through R^2. Since DW value is 2.136, we are able to evaluate that the independency of residual meets the condition. In ANOVA, investigating P-value (0.039) on F-value is suitable for the regression model 2. As the result of parameter estimates, VIF represents that multi-collinearity is stable between 1 and 3.

Table 17 The Result of Model 2 (VOLKSWAGEN)

Model	R^2	Adjusted R^2	Durbin-Watson	ANOVA		Variables	Unstrdzed. Coefficient		Strdzd. Coeffi.	t	sig.	Collinearity Stat.
				F	P		B	Std. Error	B			VIF
2	0.551	0.304	2.136	1.743	0.039	(Constant)	0.039	0.013		3.076	0.005	
						Sadness	0.013	0.032	0.068	0.408	0.286	1.105
						Anger	-0.166	0.185	**-0.214**	-2.893	**0.038**	2.316
						Fear	-0.041	0.061	-0.122	-0.673	0.207	1.331
						Delight	-0.004	0.090	-0.012	-0.045	0.165	2.957
						Satisfaction	0.072	0.069	0.255	1.031	0.311	2.464
						Shame	-0.023	0.049	-0.133	-0.458	0.150	3.366
						Frustration	-0.169	0.086	**-0.504**	-1.980	**0.058**	3.739

As the result of significancy of each variable, it assumes different aspect from the result of model 1. P-value of domains Anger (0.038) and Frustration (0.058) have significant influence on the VOLKSWAGEN's market share. We focus on the result. Frustration had been evaluated that it had been divided into the factor of neutrality from PCA in Phase II. However, as seeing the standardized coefficient, Frustration (-0.504) has a negative (-) influence on market share of VOLKSWAGEN. At this point, we are able to infer that Frustration works as a negative factor on it.

In case of VOLKSWAGEN, the findings of "Dieselgate" were provided to the California Air Resources Board (CARB) in May 2014 and erupted on 18 September 2015, when the United States Environmental Protection Agency (EPA) issued a notice of violation of the Clean Air Act to German automaker VOLKSWAGEN Group after it was found that VOLKSWAGEN had intentionally programmed turbo-charged direct injection (TDI) diesel engines to activate certain emissions controls only during laboratory emissions testing. In spite of the event, the market share had risen rather than fallen from 3.3% to 4.1%. It is likely that this increasing aspect of market share is due to the sales promotion to minimize the deficit by "Dieselgate." However, it rapidly declined from 4.1% to 3,2% in December 2015. According to the result from linear regression analysis, it suggests that the negative factors of Anger and Frustration had presented in VOC and had a significant influence on its performance.

5. Conclusion

5.1 Summary of Study

In this present study, we verified that 7 sentiment domains extracted through sentiment analysis from social media have an influence on business performance. This research consists of three phases. In phase I, we constructed the sentiment lexicon and 7 sentiment domains by analyzing VOC on 26 auto companies through sentiment analysis technique.

In phase II, in order to retain the reliability of experimental data, we examined auto-correlation analysis and PCA. The findings from the auto-correlation analysis proved auto-correlation and the sequence of the sentiments. We could notice that, although there are slight differences in each respective sentiment domain, it generally has a significant amount of the auto-correlation coefficient in the sequence of 5 and 7 months. In addition, the results from PCA reported that the 7 sentiment domains are connected with positivity, negativity and neutrality. Sadness, Anger, and Fear has been divided into a principal component 1. Shame and Frustration has been divided into the principal component 2. Delight and Satisfaction are connected to the principal component 3. The principal component 1 represents the negativity on the auto market. The principal component 2 includes neutrality. The principal component 3 including Delight and Satisfaction mainly shows the positivity on the market, but the principal component 1, negative factor, explains 52.344% of the entire components. Furthermore, we focused on the fact that Shame and Frustration have not been included in any group of the negative or the positive. Thus, it contains the neutral information or is regarded as the noise regardless of the plus (+) or minus (-) on the auto sales.

In phase III, we investigated how 7 domains impact on the market share of four major auto companies such as GM, FORD, FCA, and VOLKSWAGEN by using linear regression analysis. The results indicated that the sentimental factors have a significant influence on the actual market share. In case of GM, we could reason that Sadness of sentiment domains on customer reviews had negatively impacted on the business performance on a significantly high level (-0.492 in model 1 and -0.555 in model 2), compared to the market share from January, 2013 to December, 2015(from 18.7% to 17.7%). In FORD's case, Sadness and Anger have negative (-) influence on market share, while Delight affects on it on a bit high level. In case of FCA, the market share had increased 2.1% point from 11.3% to 13.4%. Even though negativity and positivity had coexisted during the period, we are able to infer that the positive sentiment factors such as Delight and Satisfaction in VOC had positively impacted on the business performance of FCA. In case of VOLKSWAGEN, Anger and Frustration have significant influence on the VOLKSWAGEN's market share. In particular, Frustration had been evaluated that it had been divided into the component of neutrality from PCA in Phase II. However, as seeing the standardized coefficient, Frustration (-0.504) has a negative (-) influence on market share of VOLKSWAGEN. At this point, we are able to infer that Frustration works as a negative factor on it. In spite of "Dieselgate," the market share had even risen rather than fallen from 3.3% to 4.1%. It is likely that this increasing aspect of market share is due to the sales promotion to minimize the deficit by "Dieselgate." However, it was rapidly declining in December 2015, the next month. According to the result from linear regression analysis, it suggests that the negative factors of Anger and Frustration had presented in VOC and a significant influence on its performance.

5.2 Managerial Implication and Limitation

This present study can have a practical and diverse effect on actual corporate management activities. Sentiment and financial variables can be used to predict the market situation of the organization and industry and understand and forecast the trend of products and services.

In addition, by combining external VOC data with internal structured data, organizations can cope with opportunities in management. Through sentiment analysis of customer reviews, we can establish marketing strategy by analyzing consumers emotion pattern, finding hidden consumers' needs and forecasting changes in customers' behavior. Moreover, it is possible to induce reduction in complaints by grasping consumer complaints, and obtain an opportunity to improve the company-wide service and manage its risk by identifying association connected to 'branch - service - customer complaints.'

Lastly, financial companies seek for solution to prevent departure of customers such as termination and improve service to customers via VOC analysis information for customer management. The VOC information combined with the distribution data related to the logistics provides the logistics information requested by the customer and grasps the customer's needs and is utilized for the products inventory prediction. As such, this study contributes to various industries in general, and further studies on various industrial fields related to this study are needed.

The three different aspects between this study and previous studies are provided as follows.

Construction of sentiment lexicon;

In order to meet the research purpose, which is to verify the influence of emotion on the performance of the corporation, more precise analysis of the VOC of each company was promoted by building the sentiment lexicon specialized in the automobile market. Vocabulary in the lexicon used for general purpose can cause many problems in terms of accuracy and meaning for sentiment analysis. The sentiment lexicon related to automotive industry was built in order to compensate the defects. The process of lexicon can be a standardized procedure for opinion mining. Based on the sentiment lexicon, it is possible to be applied to not only analysis on automotive industry, but also research on a variety of industry fields.

Embodiment of reliability on data;

Researchers can assure reliability on data for experiment by suggesting the methods; auto-correlation analysis and PCA. We could grasp the pattern of sentiment through auto-correlation analysis and make a group of respective sentiment through PCA. It could be one way to solve the problem of data reliability in big data research.

The study of neutrality;

Generally, although researchers tend to regard neutral sentiment as useless "buzz", through the empirical experiment based on two regression models, we intended to investigate hidden elements of neutral emotion. The sentiment of neutrality had a negative influence on the automotive market share according to the results from linear regression analysis. Thus, it is necessary to study about the neutrality additionally.

As limitation of the present research, we extracted and utilized the only vocabulary in the middle of process selecting sentiment words. We need to expand the boundary of study into not just vocabulary, but the chunk of words considering context, nuances and emoticons or symbols representing various emotions. In addition, it is necessary to classify more detailed and specific emotional domains without the limitation to 7 sentiment domains.

Furthermore, although it is meaningful to study of Korean industry, the research on Korean industry has not been carried out except for movie or stock due to inadequate conditions such as Korean morpheme analysis and the absence of programs or tools related to Korean sentiment analysis. In particular, not frequency, there is a difficulty in scoring for the polarity or intensity of sentiment, Therefore, the research methodology on sentiment analysis, which is suitable for Hangul system, is urgently needed. The future research will reflect these problems.

5.3 Future Study

In this study, we validated and verified the fact that the time-series information exists in the VOC sentiment data from the result through auto-correlation analysis. However, we did not conduct a forecasting experiment under the judgment that it is impossible to predict the sales and market share of the automobile market with only sentiment information in terms of prediction. As a research issue for the future, in addition to the sensibility information, we will excavate a variety of methodologies that can be used to predict the future market trends, market share, the corporate bankruptcy and innovation by adding a variety of variables that can be leveraged in the market forecasting. As we apply the availability of data to the market, and take advantage of auto-correlation of the market-related information and the sentiment, we expect that the findings will be a huge contribution to other researches on sentiment analysis as well as actual business performances in various ways.

In addition, in this present study, we performed a sentiment analysis based on the only vocabulary. On the other hand, as a research issue for the future, we will expand the research boundary to vocabulary bundle, context, sentences, and paragraph in order to find out a hidden pattern and consumer's intention by taking advantage of various data mining methodologies. We will suggest strategic directions for the innovation that organizations seek after applying and integrating a variety of research fields from semantics in linguistics to cognitive ethology based on psychology, and advance the research agenda with a variety of sources and perspectives in order to present a solution for enterprises and consumers to the customer's needs.

References

Abdi, H. & Williams, L.J., 2010. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4), 433-459.

Agarwal, A., 2012. Secret key encryption algorithm using genetic algorithm. *IJARCSSE*, 2(4), 57-61.

Aggarwal, P., Vaidyanathan, R., Venkatesh, A., 2009. Using lexical semantic analysis to derive online brand positions: An application to retail marketing research. *Journal of Retailing*, 85(2), 145-158.

Agrawal, R., Imielinski, T., & Swami. A., 1993. Mining association rules between sets of items in large databases. *Proceedings of the ACM SIGMOD International Conference on Management of Data* (SIGMOD '93), 207-216.

Alpers, G., Winzelberg, A., Classen, C., Roberts, H., Dev, P., Koopman, C., & Taylor, C., 2005. Evaluation of computerized text analysis in an internet breast cancer support group. *Computers in Human Behavior*, 21(2), 361–376.

Alvin C. Rencher & William F. Christensen, 2012. *Methods of Multivariate Analysis, Wiley Series in Probability and Statistics*. 709 (3rd ed.), John Wiley & Sons, 19.

Andrews, R., Diederich, J., & Tickle, A.B., 1995. A survey and critique of techniques for extracting rules from trained artificial neural networks. *Knowledge-Based Systems*, 8(6), 378-389.

Ansari, S., Chetlur, S., Prabhu, S., Kini, G.N., Hegde, G., & Hyder, Y., 2013. An overview of clustering analysis techniques used in data mining. *International Journal of Emerging Technology and Advanced Engineering*, 3(12), 284–286.

Argamon, S., Koppel, M., & Avneri, G., 1998. Routing documents according to style. *First International Workshop on Innovative Information Systems*, 85-92.

Bantum, E. & Owen, J., 2009. Evaluating the validity of computerized content analysis programs for identification of emotional expression in cancer narratives. *Psychological Assessment*, 21(1), 79-88.

Bollen, J., Mao, H., & Zeng, X., 2011. Twitter mood predicts the stock market. *Journal of Computational Science*, 2, 1-8.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R., 2000. *CRISP-DM 1.0: Step-by-Step Data Mining Guide*. NCR Systems Engineering Copenhagen (USA and Denmark), Daimler Chrysler AG (Germany), SPSS Inc. (USA) and OHRA Verzekeringenen Bank Group B.V. (The Netherlands).

Cha, S., Kang, J., Choi, J., 2012. A study on social media opinion mining based enterprise crisis management. *Proceedings of KIISE Conference*, 39(1), 142-144.

Choi, Y.-J., Choi, H., 2011. A study on the customer satisfaction strategies of the online company using VOC. *Journal of Korean Industrial Economics and Business*, 3(1), 73-93.

Coussement, K. & Van den Poel, D., 2008. Improving customer complaint management by automatic email classification using linguistic style features as predictors. *Decision Support Systems*, 44(4), 870-882.

Darrah, M., Taylor, B., & Webb, M., 2005. A geometric rule extraction approach used for verification and validation of a safety critical application. *2005 Florida Artificial Intelligence Research Society Conference*, FL, U.S.A.

Diener, E., Smith, H. Fujita, F., 1995. The personality structure of affect. *Journal of Personality and Social Psychology*, 69, 130.

Duan, W., Gu, B., & Whinston, A.B., 2008. Do online reviews matter? -An empirical investigation of panel data. *Decision Support Systems*, 45(4), 1007-1016.

Esling, P. & Agon, C., 2012. Time-series data mining. *ACM Computing Surveys,* 45(1), 34.

Esuli, A. & Sebastiani. F., 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. *LREC*, 417-422.

Finch, B.J., 1999. Internet discussions as a source for consumer product customer involvement and quality information: an exploratory study. *Journal of Operations Management*, 17(5), 535-556.

Freedman, D.A., 2009. *Statistical Models: Theory and Practice.* Cambridge University Press, 26.

Fu, T.C., 2011. Review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24(1), 164–181.

Gosain, A. & Bhugra, M., 2013. A comprehensive survey of association rules on quantitative data in data mining. *Proceedings of the IEEE Conference on Information & Communication Technologies (ICT '13)*, 1003–1008.

Herbert A. Edelstein, 1999. *Introduction to Data Mining and Knowledge Discovery.* Third Edition ISBN: 1-892095-02-5, Two Crows Corporation, MD, U.S.A.

Holland, J. H., 1975. *Adaptation in Natural and Artificial Systems.* MIT Press.

Jain, A.K. & Dubes, R.C., 1988. *Algorithms for Clustering Data.* Prentice Hall, Englewood Cliffs, NJ, U.S.A.

Jesty, P.H., Hobley, K.M., Evans, R., & Kendall, I., 2000. Safety analysis of vehicle-based systems, in: F. Redmill, T. Anderson (Eds.). Lessons in System Safety. *Proceedings of the 8th Safety-Critical Systems Symposium (SCSS),* Springer, London.

Jung, 2013. The influence of negative emotions on customer contribution to organizational innovation in an online brand community. *Journal of Korean Society for Internet Information*, 14(4), 91-100.

Jung, H.W. & Nah, K., 2007. A Study on the Meaning of Sensibility and Vocabulary System for Sensibility Evaluation. *Journal of the Ergonomics Society of Korea*, 26(3), 17-25.

Kahn, J., Tobin, R., Massey, A., & Anderson, J., 2007. Measuring emotional expression with the linguistic inquiry and word count. *American Journal of Psychology*, 120(2), 263–286.

Katz, Gerald M., 2001. *One right way to gather the voice of the customer*. PDMA Visions Magazine.

Kesavaraj, G. & Sukumaran, S., 2013. A study on classification techniques in data mining. *Proceedings of the 4th International Conference on Computing, Communications and Networking Technologies (ICCCNT '13)*, 1-7.

Kessler, B., Numberg, G., & Schütze, H., 1997. Automatic detection of text genre. *Meeting of the Association for Computational Linguistics*, 32-38.

Kim, D. -Y., Yong, W., Park, H. -R., 2012. Constructing an Evaluation Set for Korean Sentiment Analysis Systems Incorporating the Category and the Strength of Sentiment. *International Journal of Contents*, 12(11), 30-38.

Kim, J.S., Jin, S., 2013. A study on the application of opinion mining based on big data. *Journal of the Korean Data Analysis Society*, 15, (1B), 101-11.

Kim Y., Jeong, S.R., & Ghani, I., 2014. Text opinion mining to analyze news for stock market prediction. *International Journal of Advance. Soft Comput. Appl*, 6(1).

Kim, Y., Kim, N., & Jeong, S.R., 2012. Stock-index invest model using news big data opinion mining. *Journal of Intelligence and Information Systems*, 18(2), 143-156.

Kozinets, R., 2002. The field behind the screen: Using netnography for marketing research in online communities. *Journal of Marketing Research*, 39(1), 61-72.

Kujawski, B., Hołyst, J., & Rodgers, G.J., 2007. Growing trees in internet news groups and forums. *Physical Review*, 76, 103.

Larose, D.T., 2005. *Discovering Knowledge in Data: An Introduction to Data Mining*. ISBN 0-471-66657-2, John Wiley & Sons, Inc.

Lee, G.H., & Lee, K.J., 2013. Twitter sentiment analysis for the recent trend extracted from the newspaper article. *KIPS Transactions on Software and Data Engineering*, 2(10), 731-738.

Lee, H., Han, J.Y., & Suh, Y., 2014. Gift or threat? An examination of voice of the customer: The case of MyStarbucksIdea.com. *Electronic Commerce Research and Applications*, 13, 205-219.

Lerner, J.S., & D. Keltner, 2000. Beyond valence: Toward a model of emotion-specific influences on judgment and choice. *Cognition & Emotion*, 14, 473-493.

Li, N., & Wu, D.D., 2010. Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decision Support Systems*, 48, 354-368.

Liu & Bing, 2010. Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2, 627-666.

Ljung, G.M., & Box, G.E.P., 1978. On a measure of a lack of fit in time series models. *Journal of the American Statistical Association*, 65, 1509–1526.

Loughran, T., & McDonald, B., 2011. When is a liability not a liability? Textual analysis dictionaries, and 10-Ks. *Journal of Finance*, 661(1), 35-65.

Mentzas, G., 2011. *Social media provides huge opportunities, but will bring huge problems*. The Economist, 50.

Miller, G.A., 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38.

Nasukawa, T., & Yi, J., 2003. Sentiment analysis: Capturing favorability using natural language processing. *The 2nd International Conference on Knowledge Capture*, 70-77.

Nayak, R., Hayward, R., & Diederich, J., 1997. Connectionist knowledge base representation by generic rules from trained feed forward neural networks. *Proceeding of Connectionist Systems for Knowledge Representation and Deduction Workshop*, Townsville, Australia.

O'Leary, D.E., 2011. Blog mining-review and extensions: from each according to his opinion. *Decision Support Systems*, 51(4), 821-830.

Pang, B., Lee, L., & Vaithyanathan, S., 2002. Thumbs up?: sentiment classification using machine learning techniques. *The ACL-02 Conference on Empirical Methods in Natural Language Processing*, 10, 79-86.

Park, K.-M., Park, H., Kim, H.-G., & Ko, H., 2013. Review mining using lexical knowledge and modality analysis. *In Proceedings of the 5th International Universal 5th International Universal Communication Symposium(IUCS)*, 54-60.

Pearson, K., 1901. *On lines and planes of closest fit to systems of points in space.* Philosophical Magazine, 2(11), 559–572.

Robert, A., Diederich, J., & Tickle, A.B., 1995. A survey and critique of techniques for extracting rules from trained artificial neural networks. *Knowledge Based Systems*, 8, 373-389.

Rokach, L. & Maimon, O., 2008. *Data Mining with Decision Trees: Theory and Applications (Series in Machine Perception and Artificial Intelligence)*. ISBN: 981-2771-719, World Scientific Publishing Company.

Romano, N. C., Donovan, C., Chen, H., & Nunamaker, J., 2008. A methodology for analyzing web-based qualitative data. *Journal of Management Information Systems*, 19(4), 213-246.

Schumaker, R.P., & Chen, H., 1999. Textual analysis of stock market prediction using breaking financial news: the AZFin text system. *ACM Transactions on Information Systems*, 27(2).

Seal, H.L., 1967. The historical development of the Gauss linear model. *Biometrika*. 54(1/2), 1-24.

Shaw, P.J.A., 2003. *Multivariate Statistics for the Environmental Sciences*. Hodder-Arnold.

Shim, K. Yang, J., 2004. High speed Korean morphological analysis based on adjacency condition check. *Korean Institute of Information Scientists and Engineers*, 31(1), 89-99.

Song, J., & Lee, S., 2011. Automatic construction of positive / negative feature-predicate dictionary for polarity classification of product reviews. *Korean Institute of Information Scientists and Engineers*, 38(3), 157-168.

Song, S., 2011. *Analysis and acceleration of data mining algorithms on high performance reconfigurable computing platforms*. Ph.D., Iowa State University.

Spangler, S. & Kreulen, J., 2008. *Mining the Talk: Unlocking the Business Value in Unstructured Information*. IBM Press.

Spertus, E., 1997. Smokey: Automatic recognition of hostile messages. The *National Conference on Artificial Intelligence*, 1058-1065.

Subramaniam, L.V., Faruquie, T.A., Ikbal, S., Godbole, S., & Mohania, M.K., 2009. Business intelligence from voice of customer. *IEEE International Conference on Data Engineering*.

Takeuchi, H., L.V. Subramaniam., T. Nasukawa, & S. Roy., 2009. Getting insights from the voices of customers: Conversation mining at a contact center. *Information Science*, 179(11), 1584-1591.

Tetlock, P.C., Saar-Tsechansky, M., & Macskassy, S., 2008. More than words: quantifying language to measure firms' fundamentals. *Journal of Finance*, 63(3), 1437-1467.

Thelwall, M., Buckley, K., & Paltoglou, G., 2008. Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology*, 62, 406-418.

Tronvoll, B., 2011. Negative emotions and their effect on customer complaint behavior. *Journal of Service Management*, 22, 111-134.

Vedder, R.G., Vanecek, M.T., Guynes, C.S., & Cappel, J.J., 1999. CEO and CIO perspectives on competitive intelligence. *Communications of the ACM*, 42(8), 108-116.

Ward, D., Jesty, P.H., & Rivett, R.S., 2009. Decomposition scheme in automotive hazard analysis. *SAE International Journal of Passenger Cars-Mechanical Systems*, 2(1), 803-813.

Whitelaw, C., Garg, N., & Argamon, S., 2005. Using appraisal groups for sentiment analysis. *The 14th ACM International Conference on Information and Knowledge Management*, 625-631.

Wilson, T., Wiebe, J., & Hoffmann, P., 2009. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational linguistics*, 35, 399-433.

Woolridge, A., 2011. *Social media provides huge opportunities, but will bring huge problems*. Economist, 50.

Yan, X., & Xiao, G.S., 2009. *Linear Regression Analysis: Theory and Computing*. World Scientific, 1–2.

Yang S. & Ko, Y., 2014. Classifying Korean comparative sentences for comparison analysis. *Natural Language Engineering*, 20(4), 557-581.

Yu, E., Kim, Y., Kim, N., & Jeong, S., 2013. Predicting the direction of the stock index by using a domain-specific sentiment dictionary. *Journal of Intelligence and Information Systems*, 19(1), 95-110.

Yune, H., Kim, H.-J., & Chang, J.-Y., 2010. An efficient search method of product review using opinion mining techniques. *Journal of KIISE : Computing Practices and Letters*, 16(2), 222-226.

Zhuang, L., Jing, F., Zhu, X.Y., & Zhang, L., 2008. Movie review mining and summarization. *Conference on Information and Knowledge Management: Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, 43-50.

www.cargurus.com

www.daumsoft.com

www.goodcarbadcar.net

www.TextAnalysisOnline.com

www.wikipedia.org

www.wordnet.princeton.edu

Automakers & ANDC (2013-2015) United States Auto Sales Brand Rankings.

Consumer Report (2013-2015).

U.S. News & World Report (2013-2015).