ISSN 1229-1889(Print)
ISSN 2287-9005(Online)
J Korean Soc Qual Manag Vol. 45, No.1:117-128, March 2017
http://dx.doi.org/10.7469/JKSQM.2017.45.1.117

감성분석과 Word2vec을 이용한 비정형 품질 데이터 분석

이진욱* · 유국현** · 문병민* · 배석주*†

*한양대학교 산업공학과
**한양대학교 수학과

Informal Quality Data Analysis via Sentimental analysis and Word2vec method

Chinuk Lee* · Kook Hyun Yoo** · Byeong Min Mun* · Suk Joo Bae*†

*Department of Industrial Engineering Hanyang University
**Department of Mathematics, Hanyang University

ABSTRACT

Purpose: This study analyzes automobile quality review data to develop alternative analytical method of informal data. Existing methods to analyze informal data are based mainly on the frequency of informal data, however, this research tries to use correlation information of each informal data.

Method: After sentimental analysis to acquire the user information for automobile products, three classification methods, that is, naïve Bayes, random forest, and support vector machine, were employed to accurately classify the informal user opinions with respect to automobile qualities. Additionally, Word2vec was applied to discover correlated information about informal data.

Result: As applicative results of three classification methods, random forest method shows most effective results compared to the other classification methods. Word2vec method manages to discover closest relevant data with automobile components.

Conclusion: The proposed method shows its effectiveness in terms of accuracy and sensitivity on the analysis of informal quality data, however, only two sentiments (positive or negative) can be categorized due to human errors. Further studies are required to derive more sentiments to accurately classify informal quality data. Word2vec method also shows comparative results to discover the relevance of components precisely.

Key Words: Naïve Bayes, Random Forest, Sentimental Analysis, Support Vector Machine, Text Mining, Word2vec.

• Received 6 March 2016, 1st revised 21 March 2017, accepted 22 March 2017
† Corresponding Author(sjbae@hanyang.ac.kr)
ⓒ 2017, The Korean Society for Quality Management
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0) which permits unrestricted non-Commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
※ 본 연구는 2015년도 산업통상자원부의 재원으로 한국에너지기술평가원(KETEP)의 지원을 받아 수행한 연구 과제입니다. (No. 20154030200900)

1. 서 론

소셜 네트워크 서비스, 블로그 등의 개인과 개인 간의 간접적인 상호작용 서비스가 증가함에 따라 기업의 서비스에 따른 소비자의 평가방법도 다양하게 증가하였다. SNS나 블로그등의 인터넷 활동에서 소비자의 제품 평이나 서비스에 대한 평가를 쉽게 찾을 수 있는 가운데, 비정형화된 데이터의 분석방안 또한 발전해가고 있는 추세이다. 하지만 대다수의 비정형 데이터, 텍스트 마이닝은 주로 데이터의 빈도수를 고려하는 방식을 사용한다. 대다수의 방식이 데이터가 가진 의미 및 뜻을 사용하는 대신, 얼마나 자주 사용되는 지, 혹은 다른 데이터와 얼마나 자주 연관이 되는 지 등의 빈도를 주로 고려한다. 대표적인 예로 문서나 기사의 키워드를 시각적으로 보여주는 워드 클라우드(Word cloud), 문서나 지상에서 특정 단어의 빈도에 따라 가중치를 부여하는 TF-IDF(Term frequency-Inverse document frequency) 가중치 적용 방법 등이 있다. 이러한 방법은 단어의 빈도수를 사용하여 직관적으로 주제 및 관심분야를 찾을 수 있지만, 데이터에 존재하는 소비자들의 심리 등을 심층적으로 발취하는데 한계점이 존재한다.

본 연구는 빈도수를 주로 고려하는 비정형 데이터의 대안으로써 제품에 대한 비정형 형태의 소비자 품질 평가 데이터를 감성분석(Sentimental analysis)를 적용한 뒤 순수 베이지안 분류기(Naïve Bayes), 서포트 벡터 머신(Support vector machine), 랜덤 포레스트(Random forest) 방법을 통해 분류한 결과에 대한 효과 및 정확도에 대한 비교 분석을 실시하였다. 또한 Word2vec를 사용하여 데이터에 대한 연관성 분석을 실행하였다. 이러한 분석 방법을 통해 소비자의 사용 차량의 품질평가에 대한 감성정보를 발취, 서비스나 제품에 대한 소비자의 정보를 확보할 수 있을 것으로 기대하며, 또한 Word2vec방법을 사용하여 빠른 계산을 통하여 데이터간의 상관분석을 실시, 중요 단어 추출 가능성을 검토, 분석하였다.

2. 본 론

2.1 연구배경 및 문헌 조사

소셜 네트워크 서비스, 블로그 등의 개인과 개인 간의 간접적인 상호작용 서비스가 증가함에 따라 기업의 서비스에 따른 소비자의 평가방안도 다양하게 증가하였다. SNS나 블로그 등의 인터넷 활동에서 소비자의 제품 평이나 서비스에 대한 평가를 쉽게 찾을 수 있는 가운데, 이런 비정형화된 데이터의 분석방안으로써 데이터 마이닝 및 오피니언 마이닝 기술이 발전해가고 많은 연구가 진행 중이다. 대표적인 방법으로 Lee and Kim (2009)은 TF-IDF(Term frequency-Inverse document frequency)를 사용하여 키워드 추출 방법을 적용하였으며, Tseng et al. (2009)은 TFC(Total frequency in cluster)를 사용하여 문단 전체의 키워드를 통한 주제별 분류 방법을 연구하였다. 또한 Yong et al. (2009)은 데이터의 클러스터링에 대한 KNN(K nearest neighbor) 알고리즘 적용을 통한 분석 방안을 연구하였다. 이러한 대다수의 연구 방법이 키워드의 빈도수에 집중, 비정형 데이터에 대한 소비자의 감성을 발취하기 힘든 단점이 있다. 본 연구에서는 단어의 빈도수 분석이 아닌 소비자의 감성을 기반으로 비정형 데이터인 차량의 서비스에 대한 소비자 리뷰 데이터 분석을 시도하였다. 현재 주목받고 있는 방법인 감성분석은 비정형 데이터에 적용되고 있는 방법 중 하나로써 고객의 욕구 및 감성을 진단하여 다양한 분석을 가능하게 하며 여러 산업에 적용할 수 있다. 감성분석의 적용사례로서 Lee et al. (2013)은 감성 분석 방법을 이미지 및 비정형 뉴스 데이터에 적용하여 주식시장의 주가 예측을 시도하였다. 또한 Kim et al. (2011)은 감성분석을 사용하여 투자의사 결정모형을 구축하였

다. 다른 비정형 데이터 분석 사례로써 Quoc and Mikolov (2014)와 Mikolov et al. (2003)이 데이터를 벡터 형태로 변환하여 분석하는 새로운 word2vec 방법을 제안하였다.

2.2 비정형 데이터 분류를 위한 감성분석 이론

감성분석 방법은 텍스트 마이닝(Text mining)및 오피니언 마이닝(Opinion mining)의 한 기법으로써 데이터로부터 소비자의 감성 관련 정보를 추출하는 방식이다. 주로 데이터를 작성한 사용자의 감정, 태도, 의견, 성향 같은 주관적인 데이터를 추출하고자 하는 방법으로써 컴퓨터 언어 및 자연 언어를 사용한다. 데이터의 주제나 다른 정보 추출보다는 소비자의 감성추출이 주 목적이며, 단순한 제품에 대한 소비자 의견뿐만 아니라 비정형 데이터에 들어있는 감성을 분석함으로써 보다 정확한 소비자의 정보 추출이 가능하다. 또한 감성이 들어간 대다수의 분야의 접목 및 응용이 가능하며 특히 현재의 트위터, 페이스 북 등 인터넷 매체의 발달로 인해 광범위한 소비자나 사용자의 감성을 추출할 수 있다. 이러한 단순한 인터넷 매체뿐만 아니라 뉴스 데이터 등의 비정형 데이터와 접목하여 주식시장의 주가 분석 혹은 이미지 분석으로 사용될 수 있다. 감성분석은 주로 인터넷 매체로부터 다양한 텍스트 데이터 등의 비정형 데이터를 수집한 후 주관성 탐지를 통해 감성분석에 사용될 요소만을 분리 및 분류하고, 감성과는 관련이 없는 부분 즉 주관성이 존재하지 않는 부분 및 저자의 이름 및 성별과 같은 개인정보를 걸러낸다. 이 후 이렇게 처리된 데이터에 극성 탐지(Polarity detection)을 실행한다. 이 작업을 통해 얻어진 정보로부터 긍정이나 부정적인 단어를 탐지하여 문장이나 문단의 특정 단어의 빈도수의 평균 혹은 총합을 통해 문장이나 문단의 긍정적 또는 부정 여부를 결정한다.

2.3. 순수 베이지안 분류기 (Naïve Bayes)

특성들 사이의 독립을 가정하는 베이즈 정리를 적용하는 확률 분류기의 일종으로써 문서를 여러 범주 중 하나로 판단하는 문제에 대한 대표적인 방법 중 하나이다. 순수 베이지안 분류기는 잡음 데이터가 평균 데이터에 속하게 되므로 조건부 확률을 추정하게 될 때, 고립된 잡음 데이터의 영향에 강건하며, 또한 데이터와 관련 없는 속성의 영향을 덜 받는 장점이 있다.

순수 베이지안 분류기는 클래스 데이터 Y가 주어졌을 때, 각 속성이 조건부로 독립적이라고 가정하며 각 클래스 조건부 확률을 계산한다. 아래는 속성 집합 $\boldsymbol{X}$에 대하여 d개의 속성으로 구성된 조건부 독립성 가정에 대한 수식에 해당한다.

$$p(\boldsymbol{X}|Y=y)=\prod_{i=1}^{d}p(X_i|Y=y) \qquad \text{Equation (1)}$$

다음의 식은 개N의 변수 $\boldsymbol{X}$에 대한 K개의 클래스를 가지는 Y에 대한 확률을 나타내며 이를 조건부확률을 이용한 정리는 다음과 같이 나타낼 수 있다. 또한 각 클래스에 대한 사후확률을 계산할 수 있다. Equation (1)에서 종속변수 집합 $\boldsymbol{X}$에 대한 확률은 모든 Y값에 고정되어 있으므로 분모를 최대화할 수 있는 클래스 $p(Y)\prod_{i=1}^{d}p(X_i|Y)$를 선택하면 된다. 이때, $p(Y|\boldsymbol{X})$는 사후확률, $p(Y)$는 사전확률, $p(X_i|Y)$는 우도함수에 해당한다.

$$p(Y|\boldsymbol{X}) = \frac{p(Y)\prod_{i=1}^{d} p(X_i|Y)}{p(\boldsymbol{X})} \qquad \text{Equation (2)}$$

이 때 사후 확률을 최대화 하는 최종 예측 값 Y_{map}을 Equation (3)과 같이 도출할 수 있다.

$$Y_{map} = \arg\max p(Y)\prod_{i=1}^{d} p(X_i|Y) \qquad \text{Equation (3)}$$

2.4. 랜덤 포레스트(Random forest)

랜덤 포레스트(Random forest)는 의사결정나무를 사용하는 알고리즘으로써 데이터를 부스트래핑(Bootstrapping)하여 여러 개의 트리로 구성된 랜덤 포레스트를 형성한 앙상블 방법을 통해 데이터의 라벨을 결정하는 방식이다. 샘플의 결과물을 의사결정나무의 입력값으로 사용하여 학습하는 방식으로 인해 서로 다른 데이터로 구축되며 이로 인한 랜덤성이 생성된다. 보통 의사결정나무의 경우, 일반화 오류(Generalization error)와 과적합(Over fitting)이 생기지만 랜덤 포레스트의 경우, 임의성에 의해 서로 다른 특성을 가진 의사결정나무의 증가를 통하여 각 의사결정나무들의 예측 값이 비상관화 될 수 있도록 한다. 이러한 임의성은 일반화 성능 향상 및 랜덤 포레스트가 노이즈 포함데이터에도 잘 대응하도록 한다. 아래의 <그림 1>은 랜덤 포레스트에 대한 개념도이다.

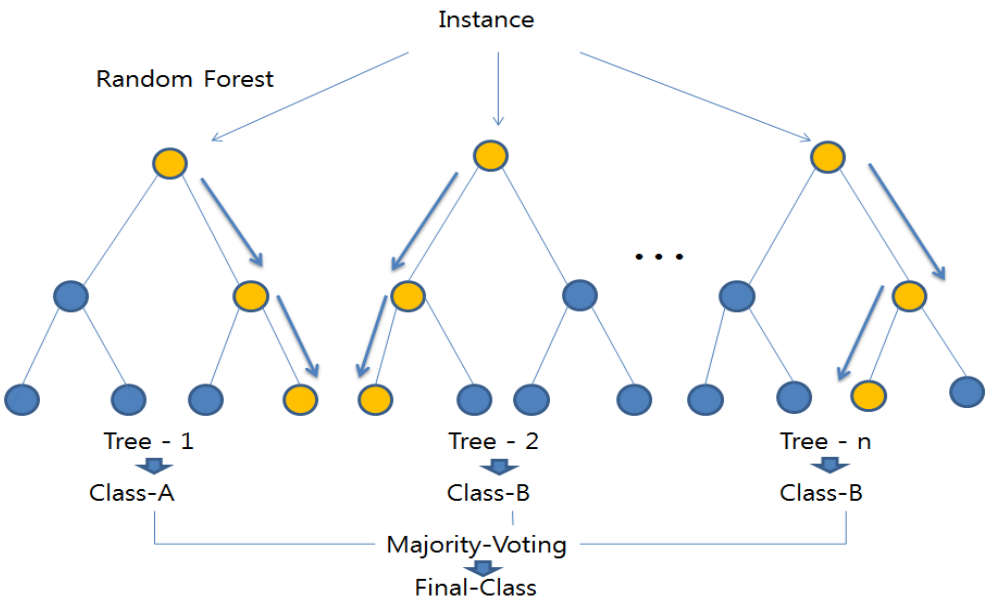

Figure 3. Random forest

랜덤 포레스트의 일반화 오류의 상한선은 다음의 식으로 나타낼 수 있다. 일반화 오류의 대한 Equation (4)을 통해서 트리 사이의 상관관계가 커지거나 앙상블의 감도가 감소하면 일반화 오류가 증가하는 경향이 존재한다. 이 때 랜덤 포레스트는 무작위성으로 인하여 각 의사결정 트리 사이의 상관관계를 줄일 수 있다.

$$\text{Generallization error} \leq \frac{\bar{\rho}(1-s^2)}{s^2} \qquad \text{Equation (4)}$$

Equation (4)에서 $\bar{\rho}$는 트리들 사이의 평균 상관관계이고 s는 트리 분류기의 세기를 측정하기 위한 수치이다. 트리 분류기의 세기는 분류기들의 평균성능을 뜻하며, 즉 분류기의 마진관계를 아래의 식과 같이 측정할 수 있다.

$$margin\ M(X,Y) = P(\hat{Y}_\theta = Y) - \max_{Z \neq Y} P(\hat{Y}_\theta = Z) \qquad \text{Equation (5)}$$

Equation (5)에서 $\hat{Y}_\theta$는 랜덤하게 정해진 벡터 θ로부터 분류기에 의해 예측된 X의 라벨로서, 마진값이 커질수록 주어진 값 X를 정확히 예측할 가능성이 증가한다(Pang, 2006, 283-287). 랜덤 포레스트의 임의화을 위한 방법은 데이터의 훈련과정에 적용되며, 주로 배깅(Bagging: bootstrap aggregating) 방법과 임의노드 최적화(Randomized node optimization)이 주로 사용된다. 배깅(Bagging)은 Bootstrap aggregating 방법의 약자로써 부트스트랩을 통해 균일한 확률분포에 따라 데이터 집합으로부터 반복적으로 샘플링을 한 훈련데이터를 기초적인 의사결정나무에 적용하여 분류기를 생성하는 방법이다. 임의노드 최적화(Randomized node optimization) 방법은 의사결정나무에서 각 각에 나누어지는 노드에 해당하는 다른 클래스 데이터들을 훈련 데이터에 이미 분류되어 있는 클래스에 해당되는 클래스로 분류하면서 적용된다. 이 방법을 통하여 다변량에 해당하는 데이터를 각 각에 해당하는 노드에 이중데이터로 분류하면서 데이터를 훈련시키는 방식이다. 또한 똑같은 형태의 의사결정나무의 생성을 없애기 위하여 각 노드를 임의적으로 생성한다.

2.5. 서포트 벡터 머신 (Support vector machine)

서포트 벡터 머신(Support vector machine)은 기계 학습(Machine learning)에서 사용되는 한 기법으로써 패턴 인식, 자료 분석을 위한 지도학급 모델이며 여러 분야에 적용되고 있는 알고리즘 중 하나이다. 주로 주어진 데이터 집합을 바탕으로 새로운 데이터가 어느 데이터 범주에 속하는지 결정하는 초평면(Hyperplane)을 가진 선형분류 모형을 사용한다. 본 문헌에서는 선형 서포트 벡터 머신을 적용하였으며 선형벡터 머신은 종종 최대 마진 분류기(Maximal margin classifier)이라 불린다. 서포트 벡터 머신이 경계를 학습하기 위해서는 먼저 선형 의사결정 경계를 정해야 한다. <그림 2>와 같이 N개의 사이즈를 가진 이진 분류문제를 가정하였을 때, 각 데이터 X에 해당하는 Y의 클래스 레이블을 표시할 수 있다. 이를 통해 선형 분류기의 의사결정 경계는 아래의 식과 같이 나타낼 수 있다. Equation (6)에서 $\boldsymbol{w}$와 b는 각각 모델의 매개 변수라 할 수 있다.

$$\boldsymbol{w} \cdot \boldsymbol{x} + b = 0 \qquad \text{Equation (6)}$$

만일 데이터 x_a와 x_b가 각 의사결정 경계에 위치한다고 하면 다음과 같은 식으로 나타낼 수 있으며 다음과 같이 추론 할 수 있다. Equation (7)에서 곱이 0이 되기 위하여 $\boldsymbol{w}$의 방향이 수직이어야 한다.

$$\begin{aligned} \boldsymbol{w} \bullet \boldsymbol{x_a} + b = 0 \\ \Rightarrow \boldsymbol{w} \bullet (\boldsymbol{x_b} - \boldsymbol{x_a}) = 0 \\ \boldsymbol{w} \bullet \boldsymbol{x_b} + b = 0 \end{aligned} \qquad \text{Equation (7)}$$

Equation (6)을 통하여 <그림 2>에 있는 의사결정 경계위에 위치한 원 데이터 ($\boldsymbol{x_s}$), 그리고 네모 데이터 ($\boldsymbol{x_c}$)에 관해서 다음과 같이 나타낼 수 있다.

$$\begin{aligned} \boldsymbol{w} \bullet \boldsymbol{x_s} + b = k \\ \Rightarrow \ k > 0,\ k' < 0 \\ \boldsymbol{w} \bullet \boldsymbol{x_c} + b = k' \end{aligned} \qquad \text{Equation (8)}$$

이 식을 통하여 원 데이터와 네모데이터의 클래스를 각기 +1과 -1로 정의한다면 가상의 데이터 U에 대한 클래스 레이블 y은 다음과 같이 추론할 수 있다.

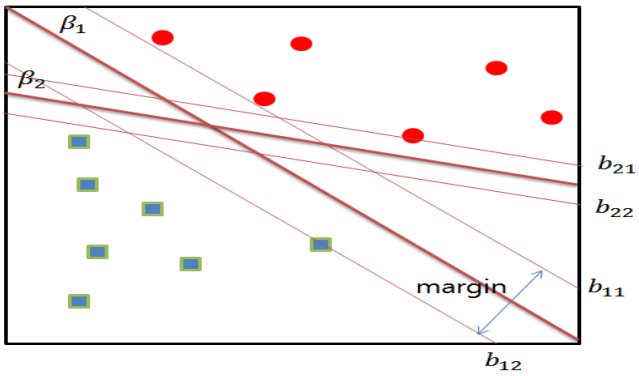

Figure 4. Support vector machine

$$y = \begin{cases} 1 \ , & \text{if } \boldsymbol{w} \bullet \boldsymbol{z} + b > 0 \\ -1 \ , & \text{if } \boldsymbol{w} \bullet \boldsymbol{z} + b < 0 \end{cases} \qquad \text{Equation (9)}$$

이렇게 클래스 레이블을 추론할 수 있는 서포트 벡터 머신의 성능을 최적화하기 위해 의사결정경계의 마진을 최대화하여야 한다. 의사결정 경계의 두 매개변수를 조정하면 두 평행한 초평면 b_{i1}과 b_{i2}에 대하여 다음과 같이 표현할 수 있다.

$$\begin{aligned} b_{i1} : \boldsymbol{w} \bullet \boldsymbol{x} + b = 1 \\ b_{i2} : \boldsymbol{w} \bullet \boldsymbol{x} + b = -1 \end{aligned} \qquad \text{Equation (10)}$$

의사결정 경계의 마진은 이 두 초평면의 거리의 의하여 다음과 같이 나타내며 b_{i1}에 위치한 데이터 점을 $\boldsymbol{x_1}$, b_{i2}에 위치한 데이터 점을 $\boldsymbol{x_2}$라 할 수 있다. 이를 적용하여 마진 d를 다음과 같이 계산할 수 있으며 이를 최대화시키므로 인해 데이터를 분류할 수 있다. (Pang, 2006, 255-256)

$$\text{Margin d} = \frac{2}{\| \text{w} \|^2} \qquad \text{Equation (11)}$$

2.5 비정형 데이터 분류를 위한 Word2vec 방법

Word2vec 방법은 문자 그대로 단어를 벡터 형태로 수치화시키는 딥러닝 방법 중 하나로 이 방법을 통해 텍스트 같은 비정형 데이터를 좌표 평면에 나타낼 수 있다. 보통 여러 단어로부터 한 단어를 추측하는 CBOW 모형(Continuous bag of words model)과 한 단어로부터 여러 단어를 추측하는 skip-gram 모형이 존재한다. Word2vec 방법 이외에도 텍스트 데이터를 벡터화하는 방식이 여러 존재하지만, Word2vec은 이러한 방식에서 좀 더 발전되어 단어의 의미를 단어가 아닌 의미 자체를 벡터형태로 표현하는 방법으로써 복잡한 개념 표현뿐만 아니라 다른 단어를 유추하는 추론까지도 구현 가능하게 하였다. 아래의 <그림 3>은 Mikolov et al. (2013)이 제안한 CBOW와 Skip-gram 모형에 대한 개념도이다.

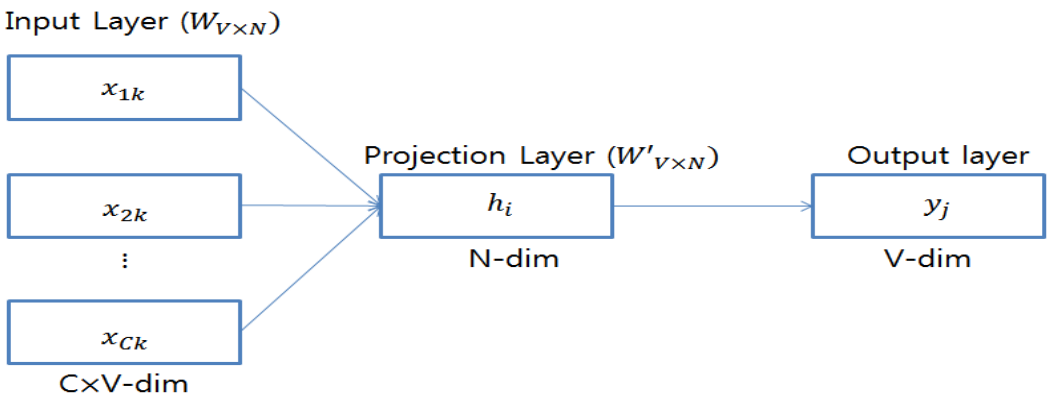

Figure 5. Continuous bag of words model

이 CBOW 모형은 입력 층(Input layer), 투사 층(Projection layer) 및 출력 층(Output layer)으로 이루어져 있으며, 입력 층에 있는 모든 단어들이 공통적으로 사용되는 $V\times N$ 크기의 투사 행렬을 통한다. 이 식의 N은 사용할 벡터의 길이를 나타내며 이 후 투사 층에서 출력 층으로 향할 때, $N\times V$ 크기의 가중치 행렬 W'를 사용한다. 입력에서 벡터화한 데이터를 투사한 후 데이터의 평균을 구해서 투사 층을 사용한다. 이 후 가중치 행렬 W'를 사용한 뒤 출력 층에 보내 구해야 할 단어를 예측한다. 아래 <그림 4>의 Skip-gram 모형도 비슷한 방식을 따르지만 CBOW

모형에서처럼 한 개의 데이터를 예측하는 방식이 아니라 한 단어로부터 연관성을 통해 여러 단어를 예측하는 방식이다.

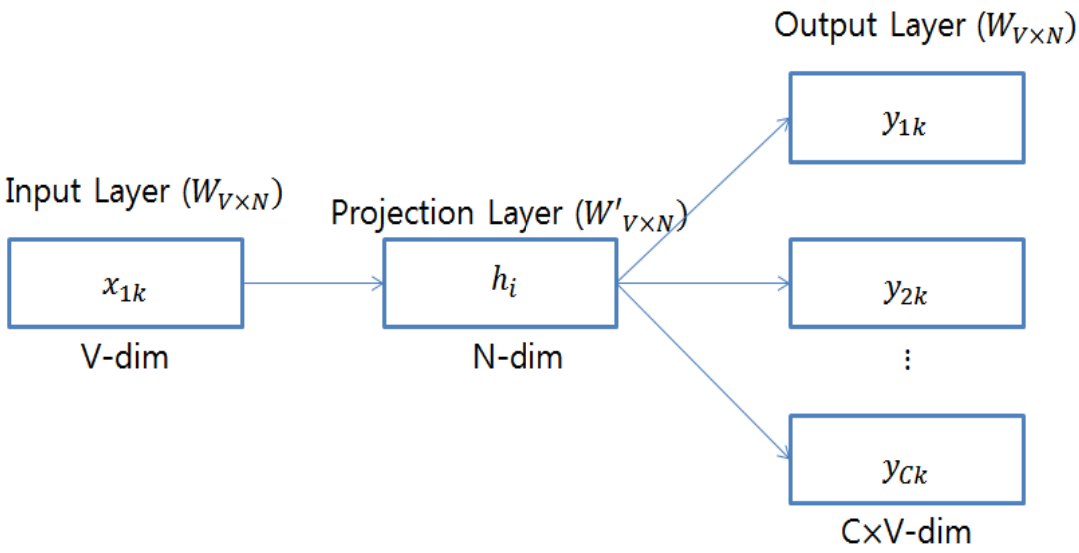

Figure 6. Skip-gram model

3. 적용 및 데이터 설명

본 연구에서는 비정형 데이터인 차량의 소비자 품질 리뷰 데이터에서 단어의 빈도수 사용 방안의 대안으로써 감성분석과 Word2vec방법을 적용하였다. H사에서 제공한 총 4,347개의 차량 품질평가 데이터에 대하여 감성분석 방법을 적용하였다. 본 품질평가 데이터는 47%의 부정적 의견, 21%의 긍정적 의견 및 32%의 중립적 의견으로 이루어져 있다. 정확한 분석을 위하여 확정적인 부정적 의견과 긍정적 의견을 제외한 데이터는 제거, 감성분석 방법을 적용, 분류한 뒤 3가지 분류 알고리즘, 즉 순수 베이지안 분류기, 랜덤 포레스트 및 서포트 벡터 머신을 적용하였다. 데이터는 training set으로써 80%, test set을 20% 할당하여 실행하였다. 랜덤 포레스트 및 서포트 벡터 머신의 경우 높은 분류정확도를 나타냈지만, 순수 베이지안 분류기의 경우 상대적으로 낮은 분류 정확도를 나타냈다. 순수 베이지안 분류기의 성능이 저조한 이유로써 베이지안 자체가 데이터의 경향을 많이 받는 경향을 보임으로 인하여 이와 같은 결과를 보여주고 있다고 추측한다.

Word2vec의 사례로써 앞서 사용한 차량 평가데이터 중 부정적 의견을 Word2vec 알고리즘에 학습하였다. Word2vec 알고리즘의 효과적인 결과 도출을 위해 parsing 방법을 통해 관사 및 전치사를 제거, 명사 위주의 불만사항을 도출하려 하였다. 그러나 Parsing 방법을 관사나 전치사 간 빈칸을 기준으로 하여 적용함으로 인하여 ‘,’ 나 ‘;’ 등의 문장기호로 연결되었을 경우, 관사나 전치사의 제거의 어려움이 있었다.

이 학습결과를 통해 부정적인 단어와 특정 명사를 연결하여 유사한 어휘를 통해 고객의 불만사항을 파악할 수 있었다. 예를 들어 만일 부정적 단어인 ‘Poor’와 가장 관련이 높은 단어로써 ‘H’회사가 연결되면, 이를 고객 불만사항으로 파악하였다. 또 다른 예시로써 ‘Navigation/Blue link’에 대하여 ‘Control’이 가장 높은 연관성을 찾으므로 각 단어에 대한 차량 부품에 대한 부정적 의견을 연결할 수 있었다.

4. 적용 결과

감성분석의 경우 중립적 의견을 제외한 총 4,347개의 고객 데이터의 68%에 해당하는 긍정 및 부정으로 분류된 데이터 중 90%를 훈련 데이터로 활용하고, 나머지 10%를 검증 데이터로 하여 3가지 방법을 적용, 분류를 실시하였다. 분류 결과 랜덤포레스트를 적용한 결과가 86.2%의 정확도로 가장 좋은 성능을 보였다. 이는 서포트 벡터 머신과 비슷한 성능을 보이지만 민감도에 관하여 약 95.9%로써 서포터 벡터 머신의 92.9%보다 향상된 성능을 보이고 있다.

Word2vec을 적용한 결과는 다음과 같이 나타내며 주요 부정적 의견의 따른 단어를 선별할 수 있었다. 이를 적용하여 부정적인 단어('poor', 'noise')와 특정명사('navigation', 'blue-link')를 결합, 고객불만 사항을 파악할 수 있었다. 이 결과를 살펴본 결과 'noise'와 가장 연관이 깊은 단어는 'door'를 추정할 수 있으며 'navigation'과 'blue-link'의 경우 'control'이 가장 연관이 높은 것으로 파악되었다. 분석 데이터가 부정적 의견에 관한 데이터이므로, 이를 기반으로 'door'에서 'noise'가 가장 많이 발생하고 고객의 불만이 일어났음을 유추할 수 있었으며, 또한 'navigation'과 'blue-link'의 'control'에 관해서 고객의 장비사용 에 대한 애로점이 파악될 수 있었다.

Table 1. Classification of sentimental analysis data and accuracy comparison

	Naïve Bayesian			Random forest			Support vector machine		
Predicted value		positive	negative		positive	negative		positive	negative
	positive	70	0	positive	44	26	positive	48	22
	negative	169	0	negative	7	162	negative	12	157
Classification accuracy(%)	29.3			86.2			85.8		

아래의 <Table 2>은 회사의 관심 분야 및 장비에 대하여 고객의 불만 데이터를 분석한 결과를 나타낸다. 단어 'Poor'의 경우, 서비스 제공 및 데이터를 제공한 'H'회사와 가장 연관이 깊은데, 이는 부정적 데이터만을 먼저 고려하여 분석한 결과이다. 다음으로 연관성 있을 단어는 'Service'와 'Dealer'라 할 수 있는데 이에 따라 'H'회사에 대한 서비스 제공에 대하여 문제 및 딜러에 대하여 문제가 제기됨을 쉽게 유추할 수 있고 이를 혼합하여 딜러의 서비스 제공의 문제점을 고려할 수 있다. 이는 부정적인 의견만을 추출하여 분석한 결과를 단어의 연관성을 고려하였으므로, 'Service'와 'Dealer'가 현대차에 대한 소비자의 부정적인 요소로 작용할 수 있을 것으로 추측 가능하다.

Table 2. Consumer complaints predicted by Word2vec

Poor	**'H'**	My	Car	Very	**Service**	About	**Dealer**	Feature	**Interior**	Before
	0.99	0.99	0.99	0.99	**0.99**	0.99	**0.98**	0.98	**0.98**	0.97
Noise	**Door**	After	**Driving**	During	Road	**Rear**	**Engine**	**Front**	Only	**Trunk**
	0.99	0.99	**0.99**	0.99	0.98	**0.98**	**0.98**	**0,98**	0.98	**0.98**
Navi gation	**Control**	**Easy**	**Not**	a	Noise	About	'H'	Only	While	**Start**
	0.99	**0.99**	**0.99**	0.99	0.99	0.99	0.99	0.99	0.98	**0.98**
Blue link	**Sounds**	**Window**	'H'	Need	**Control**	Mirror	With	Vehicle	and	Use
	0.99	**0.99**	0.98	0.98	**0.98**	0.98	0.98	0.98	0.98	0.98

5. 결론 및 향후 연구

본 연구에서는 단순히 단어의 빈도수를 사용하여 비정형 데이터를 분석하는 방안의 대안으로써 감성분석을 적용하여 분석을 실행하고자 하였다. 차량의 고객 서비스 데이터를 적용, 계산의 용이함을 위해 긍정 및 부정적인 감정으로만 데이터를 분류, 분석한 결과 높은 정확도를 보였다. 향후 연구방향에서는 이런 고객 데이터의 감성분석 결과를 좀 더 세분화된 방향, 즉 부정적 표현인 분노, 슬픔 등과 긍정적 표현인 만족, 행복 등으로 더욱 세분화하여 분석을 실시하고자 한다. 또 다른 적용 방법인 Word2vec의 경우 단어의 연관성을 쉽게 찾을 수 있는 가능성을 검토하였으며, 이를 통해 주요 부품인 Navigation이나 Blue link에 대한 문제점을 도출할 수 있었다. 이러한 방법을 통해 고객들의 제품에 대한 품질 불만사항을 쉽게 도출할 수 있을 것으로 기대한다. Word2vec을 사용하여 단어의 연관성을 알아보고자 하였을 때, 부정적인 데이터만의 분석을 고려하였고, 또한 전체 단어수가 100,000개 이상을 고려함으로써 단어 간의 연관성이 큰 차이를 보이지 않는 단점을 보였다.

이러한 단점을 보완하기 위하여 Word2vec 방법에는 좀 더 많은 단위의 비정형 데이터를 학습시키고자 하며, 데이터 확보 방안이나 다른 데이터와의 혼합을 통한 응용방안을 고려하고자 한다. 또한 부정적 데이터 분석만이 아니라 긍정적 데이터 분석을 동시 진행하여 단어 간의 연관성에 대한 자세한 분석을 실행하고자 한다.

Acknowledgement

본 연구는 2015년도 산업통상자원부의 재원으로 한국에너지기술평가원(KETEP)의 지원을 받아 수행한 연구 과제입니다. (No. 20154030200900)

REFERENCES

Eun Ji Yu, Yoo Sin Kim, Nam Gyu Kim, and Seung Ryul Jeong. 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.

Pang Ning Tang, Michael Stenbach, and Vipin Kumar. 2006. Introduction To Data Mining. Addison-Wesley Longman Publishing Co., Inc.

Quoc Le, Tomas Mikolov. 2014. “Distributed representations of Sentences and Documents.” Proceedings of the 31st international conference on machine learning, 1188-1136.

Sung-Jick Lee, and Han-Joon Kim. 2009. “Keyword extraction from news corpus using modified TF-IDF.” The Journal of Society for e-Business Studies 14(4):59-73.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2003. “Efficient estimation of word representations in vector space.” Proceedings in International Conference on learning representations 2013.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. “Distributed representation of words and phrases and their compositionality.” Proceedings in International conference on neural information processing systems, 3111-3119.

Yoo Sin Kim, Nam Gyu Kim, and Seung Ryul Jeong. 2011. “Stock-index invest model using news big data opinion mining.” Journal of Intelligence and Information Systems. Volume 18(2):143-156.

Yuen-Hsien Tseng, Chi-Jen Lin, and Yu-I Lin. 2007. “Text mining techniques for patent analysis." Information processing and management 43(5):1216-1247.

Yean Ran Lee, Eun Ju Yoon, Jung Ah Im, Young Hwan Lim, and Jung Hwan Sung. 2013. “Emotional tree using sensitivity image analysis algorithm.” Journal of the Korea Contents Association 13(11):562-570.

Zhou Yong, Li Youwen, and Xia Shixiong. 2009. “An improved KNN text classification algorithm based on clustering.” The Journal of Computers 4(3):230-237.