SNS 감성분석을 이용한 정보 추출 방법론에 관한 연구

Study on the Methodology for Extracting Information from SNS Using a Sentiment Analysis

홍 두 표* · 정 하 림** · 박 상 민*** · 한 음**** · 김 홍 회***** · 윤 일 수******

* 주저자 : 한국도로공사 전남지역본부 본부장
** 공저자 : 아주대학교 건설교통공학과 석사과정
*** 공저자 : 아주대학교 건설교통공학과 박사과정
**** 공저자 : 도로교통공단 교통과학연구원 연구원
***** 공저자 : 일마일주식회사 수석연구원
****** 교신저자 : 아주대학교 교통시스템공학과 부교수

Doopyo Hong* · Harim Jeong** · Sangmin Park** · Eum Han*** · Honghoi Kim**** · Ilsoo Yun*****

* Korea Expressway Corporation Gwangju Jeonnam Regional Headquarters
** Dept. of Civil and Transportation Eng., Ajou University
*** Traffic Science Institute, Korea Road Traffic Authority
**** Ilmile Corp.
***** Dept. of Transportation System Eng., Ajou University

† Corresponding author : Ilsoo Yun, ilsooyun@ajou.ac.kr

Vol.16 No.6(2017)
December, 2017
pp.141~155

ISSN 1738-0774(Print)
ISSN 2384-1729(On-line)
https://doi.org/10.12815/kits.2017.16.6.141

Received 3 January 2017
Revised 6 February 2017
Accepted 6 November 2017

요 약

최근 SNS 이용이 활발해짐에 따라 많은 사람들이 특정 이벤트 등에 대한 자신들의 생각을 비정형 데이터인 텍스트 형태로 자신의 SNS에 게시하고 있다. 이에 따라 금융, 유통 등 다양한 분야에서 이미 SNS를 이용하여 서비스 만족도 조사, 소비자 요구사항 모니터링, 대선 후보 선호도 등을 수행하고 있다. 하지만 교통 분야에서는 감성분석과 같은 비정형 데이터 분석을 활용하는 사례가 부족한 실정이다. 이에 본 연구에서는 한국도로공사에서 수집한 비정형 데이터인 고속도로 VOC 데이터를 이용하여 교통분야에서 사용할 수 있는 감성분석 방법론을 개발하였다. 개발된 감성분석 방법론은 수집된 비정형 데이터에 대한 형태소 분석, 감성사전 구축, 감성 판별 등으로 구성되어 있다. 개발된 방법론은 고속도로 관련 트윗 데이터를 이용하여 검증하였다. 분석 결과, 분석 기간 동안 고속도로와 관련하여 공사, 사고에 대한 정보 전달이 많이 이루어졌음을 짐작할 수 있었다. 또한 공사 및 사고로 인해 발생한 지체에 대하여 이용자들의 불만이 높았던 것으로 판단된다. 결론적으로 SNS 감성분석이 교통분야에서도 의미 있는 정보추출이 가능한 기법임을 확인하였다.

핵심어 : 감성분석, 고객의 소리, 소셜 네트워크 서비스, 트위터

ABSTRACT

As the use of SNS becomes more active, many people are posting their thoughts about specific events in their SNS in the form of text. As a result, SNS is used in various fields such as finance and distribution to conduct service satisfaction surveys and consumer monitoring. However, in the transportation area, there are not enough cases to utilize unstructured data analysis such as emotional analysis. In this study, we developed an emotional analysis methodology that can be used in transportation by using highway VOC data, which is atypical data collected by Korea Expressway Corporation. The developed methodology consists of morpheme analysis, emotional

dictionary construction, and emotional discrimination of the collected unstructured data. The developed methodology was verified using highway related tweet data. As a result of the analysis, it can be guessed that many information and information about the construction and the accident were related to the highway during the analysis period. Also, it seems that users complain about the delay caused by construction and accident.

Key words : sentiment analysis, voice of customer, social network service, twitter

Ⅰ. 서 론

1. 연구의 배경 및 목적

고속도로 관리기관인 한국도로공사는 고속도로 이용자들에게 교통 정보 제공, 긴급 구조 서비스, 휴식 시설 제공, 고속도로 관리 등 다양한 서비스를 제공하고 있다. 한국도로공사에서는 제공하고 있는 서비스의 품질관리를 위해 고속도로 이용자들을 대상으로 만족도 설문을 시행하고 있다. 하지만 전통적인 조사 방법인 설문조사는 많은 인력과 비용이 소모되기 때문에 만족도 조사와 이용자 모니터링을 위한 새로운 방법이 필요한 상황이다(Hong, 2016).

행정자치부의 2015년 주민등록 인구 통계에 따르면 국내 총 인구는 51,529,338명이다. 동일기간 스마트폰 가입자 수는 43,058,008명으로 국내 총 인구 대비 스마트폰 가입자 수는 약 84%에 해당한다. 2013년 전 세계 48개국의 모바일 이용자에 대한 조사 결과가 수록된 Google Our Planet의 보고서에 따르면 한국 이용자의 60%가 매일 스마트폰을 이용하며, 하루에 한번 이상 소셜 네트워크 서비스(Social Network Service 이하 SNS)를 이용하는 것으로 나타났다(Google, 2013). 많은 이용자들의 의견이 담긴 SNS를 이용한다면 전통적인 설문조사에 비해 쉽게 이용자의 서비스 만족도 등을 파악할 수 있을 것이다.

감성분석(Sentiment Analysis)은 텍스트 마이닝(Text Mining)의 기법으로 비정형 데이터인(unstructured data) 문서나 문장에서 작성한 사람의 감정을 추출해 내는 기술이다. 이는 문서의 주제가 아닌 해당 문서가 어떠한 감정을 가지고 있는지에 대해 분석이 가능하다(Naver, 2016). 감성 분석을 이용한 사례로 'IBK 기업은행'은 인터넷과 소셜 미디어상의 문장과 문서를 이용하여 은행의 이미지, 활동 등에 대한 감성분석을 실시하고, 이를 마케팅 및 은행 평판 관리에 활용하였다(NH Economic Research Institute, 2013).

하지만, 교통분야에서는 비정형 데이터 분석기법인 감성분석을 보편적으로 사용하지 않고 있다.

따라서 본 연구에서는 고속도로 고객의 소리(Voice of Customer, VOC) 데이터를 이용하여 감성분석을 실시하여 의미 있는 정보를 추출하고자 한다. 또한 감성분석 결과를 SNS 중 하나인 트위터(Twitter)에서 생성된 트윗(tweet) 데이터를 이용하여 의미 있는 정보를 추출된 정보가 의미 있는지를 검증하고자 한다.

2. 연구의 범위 및 절차

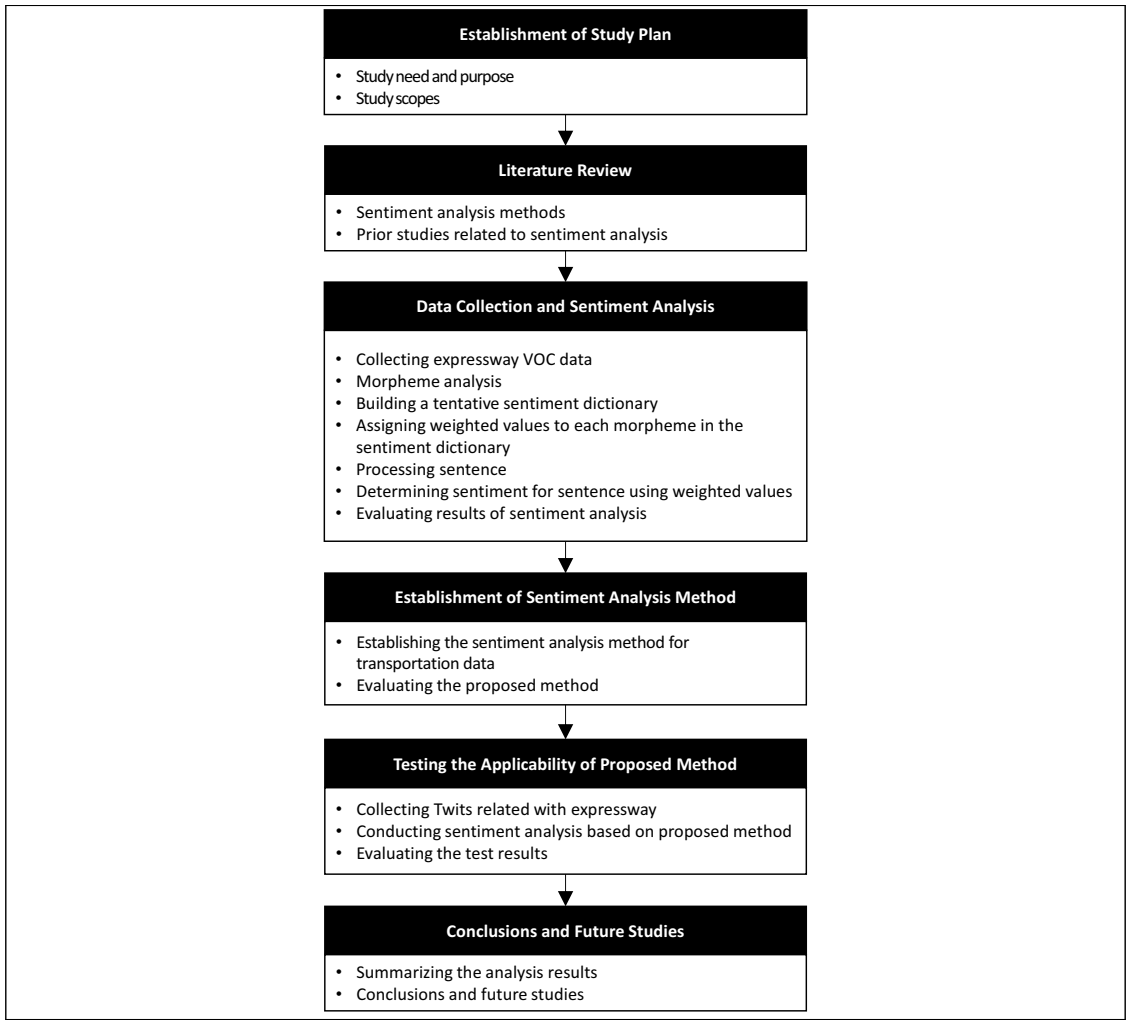
1) 연구의 범위

본 연구에서는 고속도로 VOC 데이터와 트위터 메시지인 트윗 데이터를 이용하여 감성분석을 실시하였다. VOC 데이터의 경우 2014~2015년 전체 데이터와 2016년 일부 데이터를 사용하였다. 또한 SNS의 경우 국내에서 서비스 중인 SNS 가운데 트위터를 자료수집에 사용하였다. 트위터는 안정적인 API(Application Program Interface)를 제공하여 자료 수집에 용이하기 때문이다. 2014년 5월 4일부터 6월 9일까지 약 1개월 동안의 고속도로 관련 트윗을 사용하였다.

2) 연구의 절차

교통 관련 SNS 데이터를 대상으로 감성분석을 통해 의미 있는 정보를 추출할 수 있는지에 대해 알아보고자 VOC 데이터를 이용한 감성분석 방법론을 선정하였다. 선정된 방법론이 적절한지 검증하기 위해 트윗 데이터를 이용하였으며 다음과 같은 절차에 따라 연구를 수행하였다.

우선 연구의 범위와 내용을 정한 후, 관련 이론과 기존 연구에 대한 고찰을 수행 하였다. 이 과정을 통해 감성분석에는 단어의 범용적인 의미를 사용하는 방법과 주제별 의미를 따로 사용하는 방법이 있음을 확인했다. 다음으로 한국도로공사 VOC 데이터를 수집하였다. 수집된 데이터를 이용하여 교통과 관련된 문장 및 문서에 앞의 두 가지 방법론 중 어떤 것이 적절한지 알아보았다. 감성분석을 수행하기 위해 형태소 분석, 감성사전 구축 등의 작업을 하였다. 이 과정을 통해 감성단어에 대한 가중치 부여, 임계값 설정 등을 하였고 각 데이터들에 대한 감성을 분석하였다. 그리고 분석된 결과를 바탕으로 교통부문 데이터 분석에 적절한 감성분석 방법론을 선정하였다. 앞의 과정을 통해 선정된 방법론이 적절한지 확인하기 위해 고속도로 관련 트윗을 이용하여 감성분석을 수행하였고 그 결과를 평가하였다.



〈Fig. 1〉 Study Process

Ⅱ. 관련 이론 및 연구 고찰

1. 감성분석의 정의

오피니언 마이닝(opinion mining)으로도 불리는 감성분석은 텍스트에 나타난 사람들의 의견, 태도, 성향과 같은 주관적인 데이터를 분석하는 자연어 처리 기술이다. 특히 제품 및 서비스에 대한 사용자의 의견조사, 업체에 대한 부정적 의견이나 이슈의 실시간 모니터링, 시장 현황이나 경쟁 업체의 활동, 업체의 활동이나 관련 이슈에 대한 대중의 반응 측정 등을 위해 사용되고 있다(IDG, 2014).

감성분석을 수행하기 위해서는 수집된 비정형 데이터에 대한 형태소 분석, 감성사전 구축, 감성 판별이 수행되어야 한다.

1) 형태소 분석

감성분석은 문서 혹은 문장 내에 존재하는 감성을 가진 단어들을 통해 문서 또는 문장의 긍정 및 부정을 판별하게 된다. 따라서 분석 대상이 되는 문서 또는 문장을 형태소(morpheme) 단위로 분류하여야 한다. 여기서 형태소란 의미를 가진 가장 작은 단위를 말한다. 예를 들면 영어 단어인 “users”는 “user”라는 형태소와 “-s”라는 형태소로 구분될 수 있다. 하지만, 한글과 같은 표음문자는 형태소 분석이 상대적으로 어려운 것으로 알려져 있다. 한글의 경우에는 한글 형태소 분석기를 사용하여 형태소를 분석할 수 있으며, 서울대학교의 꼬꼬마 한글 형태소 분석기와 카이스트의 한나눔 형태소 분석기가 대표적인 분석기이다.

수집된 비정형 데이터인 문장 또는 문서에 형태소 분석기를 적용하면, 문서 내에 형태소가 모두 분리가 되고, 각 형태소에 일반 명사, 수사, 대명사, 동사, 형용사, 체언, 용언, 관형사, 부사, 감탄사, 조사, 어근 등 품사를 포함한 14개의 대분류 결과가 기록된다.

2) 감성 사전

감성 사전이란 긍정, 부정과 같은 감성을 포함한 형태소들을 모아 놓은 집합체이다. 분석 대상이 되는 문장의 긍정 및 부정을 판별하기 위해서 앞서 말한바와 같이 문장의 형태소 분석이 선행되어야 하며 감성을 가지는 형태소를 제외한 나머지 형태소들을 삭제하는 문장 처리단계가 필요하다. 감성 사전은 이 단계에서 문장을 구성하는 각 형태소의 삭제 및 유지 여부를 판단하기 위한 도구로 사용된다. 따라서 감성 분석을 위해서는 감성 사전 구축이 필요하다.

감성 사전을 구축하기 위해서는 분석 문장들 내에 존재하는 감성을 포함하는 형태소들을 1차적으로 수집하게 된다. 다음으로, 감성을 표현하는데 보편적으로 사용되는 형태소만을 남기는 필터링 과정이 뒤따르게 된다. 모든 과정을 거친 후에 남아있는 형태소들이 최종적으로 감성 사전에 등재되게 된다.

3) 감성 판별

감성분석을 수행하기 위해서는 감성 사전에 등록된 형태소별 긍정 혹은 부정을 나타내는 가중치 정보가 필요하다. 감성 판별은 가중치 산출 방법에 따라 크게 범용 적인 의미에 따라 가중치를 부여하는 방법과 주제별 가중치를 계산하는 방법으로 나눌 수 있다.

2. 관련 연구 고찰

Yun et al.(2016)는 부산지역의 사회적 이슈인 해수담수화 플랜트사업에 대한 대중의 분산된 의견을 구체화 하는 것을 목적으로 하였다. 이를 위해 인터넷 카페, 블로그 게시글과 댓글, 트윗에 대한 감성분석을 실시하였다. 분석 결과 부정적 의미를 지닌 단어가 88%를 차지하였고 “방사능”, “반대” 등의 단어가 많았으며 이를 통해 사업에 대한 반대 의견뿐만 아니라 관련 시설에 대한 우려도 많음을 확인하였다.

Cho et al.(2013)는 사회적 이슈와 정치적 이슈에 대한 뉴스 댓글을 통해 댓글에 나타난 감성 단어를 기반으로 대중들의 의견을 추출하였다. 분석을 위해 네이트 뉴스의 2013년 1월 13일부터 2월 16일 까지 약 1개월 동안 ‘박근혜’, ‘박 대통령’, ‘朴 ’이 들어간 뉴스들의 댓글을 이용하였고 댓글에 나타나는 비속어와 은어들을 고려한 감성분석을 통해 실제 리서치 조사와 유사한 대중의 의견을 파악 할 수 있음을 보였다.

Bae et al.(2013)는 한국 대통령 선거 관련 트위터 분석을 통해 트위터 상에서 실시간으로 생성되는 데이터의 수집 방안을 정의하고 수집된 데이터의 분석을 통해 새로운 정보 추출이 가능함을 보였다.

Yu et al.(2013)는 주식관련 뉴스에 대한 감성분석을 통해 주가지수 등락을 예측하는 주식 도메인에 특화된 주제지향 감성사전을 구축하고 활용하는 방안을 제시하였고 범용 사전과의 비교를 통해 주제별 감성사전이 바람직함을 주장하였다.

Zhuang et al.(2006)는 다양한 장르의 영화에 대한 영화 후기에 대해 감성분석을 실시하여 긍/부정을 평가하였다. 분석 결과 낮은 수준의 긍/부정 판단 결과를 보였고 정확성 향상을 위해 실제 감성을 혼동시키는 문장에 대한 처리의 중요성을 주장하였다.

Godbole et al.(2007)는 블로그와 신문에는 최근 이슈가 되는 이벤트에 대한 보도와 함께 의견 표출이 함께 있음을 주장하였다. 이를 기반으로 블로그와 신문 상에 나타난 의견을 추출하기 위한 글 뭉치를 구성하는 단어들에 대하여 감성을 나타내는 점수를 부여하여 평가하는 시스템을 제안하였다.

3. 시사점

관련 이론 및 연구 고찰을 통해 감성분석 방법론은 감성 사전 구축 시 각 단어별 가중치 산출방법에 따라 크게 두 가지로 나눌 수 있음을 확인하였다. 따라서 두 가지 방법 중 교통분야의 비정형 데이터에 적용 가능한 적절한 방법을 찾는 것이 필요하다. 또한 교통분야에서는 아직까지 감성분석이 적용된 사례가 거의 없는 것으로 조사되었다. 따라서 본 연구를 통해 감성분석 방법론을 명확히 설정하여 기존에 사용되지 못한 교통분야 비정형 데이터를 활용할 수 있는 구체적인 분석 방법론 정립이 필요하다.

Ⅲ. 고속도로 VOC 자료 수집 및 감성분석

1. 고속도로 VOC 수집

앞에서 말한 것과 같이 감성분석을 수행하는 방법론은 감성 사전 구축 시에 해당 단어들에 범용적인 가중치를 부여하는 방법과 특정 주제에 대해 가중치를 새롭게 부여하는 방법이 있다. 교통분야에 적합한 감성분석 방법론을 알아보기 위해 사람들의 감성이 뚜렷하게 드러나는 한국도로공사 VOC 데이터를 이용하여 감성분석을 실시하였다. 분석에 사용된 자료는 2014~2015년 교통과 관련된 전체 VOC 데이터와 2016년 교통과

관련된 일부 VOC 데이터를 사용하였다. 2014~2015년 데이터는 총 1,031개로 VOC 데이터로 감성 판단에 사용되는 사전 구축과 사전 내 단어들의 가중치를 산출하기 위한 실험 데이터로 사용되었다. 2016년 일부 데이터 138개는 앞의 실험데이터를 통해 구축한 사전과 가중치의 성능을 검증하기 위한 데이터로 사용하였다.

2. 감성분석

1) 형태소 분석 및 감성사전 구축

감성분석을 위해 수집된 VOC 데이터에 대해 형태소 분석을 실시하였다. 형태소 분석에는 서울대학교 IDS(Intelligent Data Systems) 연구실에서 제공하는 “꼬꼬마 한글 형태소 분석기”를 사용하였다. 꼬꼬마 한글 형태소 분석기는 Java 라이브러리 형태로 해당 홈페이지를 통해 배포되고 있다(Seoul National University IDS Lab, 2016).

형태소 분석기를 거친 문장은 <Fig. 3>과 같이 형태소 단위로 나뉘게 된다. 이 중 감성을 포함할 수 있는 일반명사(NNG), 형용사(VA), 어근(XR)을 이용하여 감성 사전을 구축하였다. <Fig. 2>와 <Fig. 3>은 수집된 VOC 데이터와 형태소 분석을 통해 분리된 결과를 보여주고 있다.

민원내용
동일날 수원영업소를 16:15분 정상통과 한국도로공사 계측기에 하자가 있으니 과적을 불인정 하며, 과적 적발 취소 요구

〈Fig. 2〉 A Data before Morpheme Analysis

동일날	NNG(일반명사)	있	VV(동사)
수원영업소	NNG(일반명사)	으니	ECD(연결어미)
를	JKO(목적격조사)	과적	NNG(일반명사)
분	NNB (일반의존명사)	을	JX(보조사)
정상통과	NNG(일반명사)	불인정	NNG(일반명사)
한국도로공사	NNG(일반명사)	하며	JC(접속조사)
계측기	NNG(일반명사)	적발	NNG(일반명사)
에	JKM (부사격조사)	취소	NNG(일반명사)
하자	NNG(일반명사)	요구	NNG(일반명사)

〈Fig. 3〉 A Data after Morpheme Analysis

이렇게 형태소 분석을 통해 선별된 형태소 집합을 감성 사전이라 한다. 위의 예시에서 감성 사전에 속하는 형태소는 부정적인 감성을 가지는 불인정, 취소, 요구가 해당된다.

2) 함축적인 단어 및 유사단어 감성사전 구축

VOC 자료뿐만 아니라 SNS, 댓글 등에서는 문장을 줄여 쓰는 등 함축적인 단어의 사용이 높다. 따라서 교통분야 감성분석에서는 주로 사용되는 약어(예, 도로의 구조·시설 기준에 관한 규칙을 “도구시”로 칭함), 은어(예, 도로 위 불법주차나 교통법규 위반 행위 등을 “길빵”이라 칭함), 또는 사투리 등 오기가 잦은 단어(예, “휴게소”를 “휴계소”라고 칭함)들을 사전에 등록시켜 형태소 분석 과정에서 누락되지 않도록 하는 것이 중요하다. 이러한 점은 유사단어(예, 서울외곽순환고속도로, 서울외곽선, 외곽순환고속도로, 서울외곽 등)에도 동일하게 적용된다. 따라서 이러한 함축적인 단어 및 유사단어에 대한 감성사전은 처음 구축한 후에 새로운 단어가 확인되면 지속적으로 추가하는 것이 필요하다.

3) 감성 사전 내 형태소별 가중치 부여

다음으로 앞의 과정을 통해 구축된 감성 사전 내 단어들에 가중치를 부여하였다. 가중치를 부여하는 방법

은 앞서 말한 것과 같이 두 가지 방법으로 진행하였다. 첫 번째로 범용적인 의미를 사용하는 방법인 General Meaning이다. 이 방법은 부정을 의미하는 단어에는 -1, 긍정을 의미하는 단어에는 +1 값을 분석가가 주관적으로 부여하였다. 이 방법의 경우 각 단어의 가중치를 쉽게 구할 수 있는 장점이 있으나 특정 주제에 대한 감성 판단 능력이 떨어진다는 단점이 있다. 두 번째 방법은 특정 주제에 대한 가중치를 부여 하는 방법으로 Subject Meaning이다. 이 방법의 경우 <Fig. 4>와 같은 방법을 이용하여 각 단어에 가중치를 부여하였다. 주제별 단어 가중치를 구할 경우 해당 주제에 대한 감성 판단 정확도가 높은 장점이 있지만 범용적인 의미를 사용하는 방법에 비해 가중치 산출 단계가 복잡하다는 단점이 있다.

$$Term(i,j) = \begin{cases} 1 \begin{pmatrix} if\ Doc(j)\ include\ Term(i) \\ and\ Doc(j)\ is\ positive \end{pmatrix} \\ -1 \begin{pmatrix} if\ Doc(j)\ include\ Term(i) \\ and\ Doc(j)\ is\ negative \end{pmatrix} \end{cases}$$

$$Term(i).NmDocs = \#\ of\ \ VOC\ including\ Term(i)$$

$$Term(i).Opinion = \frac{\sum_{i=1}^{n} Term(i,j)}{Term(i).NmDocs}$$

〈Fig. 4〉 Formula of Subject Meaning Method (source : Kim et al., 2013)

앞의 두 가지 방법으로 단어별 가중치를 계산한 결과는 아래 <Table 1>과 같다.

〈Table 1〉 Result of Assigned Weight by two Method

Word	General Meaning	Subject Meaning	Word	General Meaning	Subject Meaning
	Weight	Weight		Weight	Weight
“감사”	1	0.0946	“불만제의”	-1	-1
“불친절”	-1	-0.7143	“지연”	-1	-0.9
“분통”	-1	-1	“하소연”	-1	-1
“책임전가”	-1	-1	“못하”	-1	-0.7083
“불쾌감”	-1	-1	“억울”	-1	-0.9355
“속상하”	-1	-1	“적재불량”	-1	-0.8333
“욕설”	-1	-1	“불법적재물”	-1	-1
“다툼”	-1	-0.6667			

4) 문장 처리

감성 사전 구축과 가중치 부여를 완료한 후에 형태소 단위로 분리된 VOC 문장들을 감성 사전에 등재된 단어들만 남기고 삭제하는 작업이 필요하다. 이 과정을 거치게 되면 각 문장은 <Table 2>와 같은 형태가 된다. 앞의 과정들을 통해 사전 내 단어들로 이루어진 형태소 단위의 VOC 데이터를 얻었고 사전 내 단어들을 이용하여 가중치를 산출하였다.

이 두 가지 자료를 이용하여 형태소 단위로 분리된 VOC에 각 단어에 해당하는 가중치를 입력하였으며, 그 결과는 각각 <Table 3>와 <Table 4>와 같다.

〈Table 2〉 VOC Data processed by Sentiment Dictionary

No.	Extracted morpheme				
1	“안되”	“억울”	“없다”	“억울”	
2	“은인”	“사고”	“감사”	“당황”	“배려”
3	“안되”	“개선”	“개선”	“거짓말”	“정체”
4	“괜찮”				
5	“사고”	“좋”	“좋”	“사고”	“사과”
6	“빠르”	“빠르”	“정체”	“빠르”	
7	“지나치”	“위험”	“아찔”	“감사”	
8	“도움”				
9	“잘못”	“걱정”	“안심”	“고맙”	“감사”
10	“죄송”	“사고”	“좋”	“짜증”	

〈Table 3〉 Assigned Weights using General Meaning Method

No.	Extracted morpheme				
1	-1	-1	-1	-1	
2	1	-1	1	-1	1
3	-1	-1	-1	-1	-1
4	1				
5	-1	1	1	-1	-1
6	1	1	-1	1	-
7	-1	-1	-1	1	-
8	1				
9	-1	-1	1	1	1
10	-1	-1	1	-1	

〈Table 4〉 Assigned Weights using Subject Meaning Method

No.	Extracted morpheme				
1	-0.7045	-0.9355	-0.8261	-0.9355	
2	1	-0.6198	0.0946	0.3125	-0.2500
3	-0.7045	-0.9281	-0.9281	-1	-0.9064
4	-0.1111				
5	-0.6198	-0.7297	-0.7297	-0.6198	-0.7333
6	-0.4286	-0.4286	-0.9064	-0.4286	
7	-0.3333	-0.4929	-0.125	0.0946	
8	0.2034				
9	-0.8793	0.3333	0.6364	0.3438	0.0946
10	-0.3077	-0.6198	-0.7297	-0.7143	

5) 가중치를 이용한 감성판단

각 문장의 감성을 판단하기 위해서 방법별로 기준을 세웠다. 범용적 의미를 사용한 방법의 경우 각 문장의 감성을 판단하는 임계값을 0으로 정하였으며 가중치들의 합이 0보다 클 경우 긍정, 작을 경우 부정, 같은 경우 중립으로 판단하였다. 주제별 의미를 사용하는 방법은 각 주제에 맞는 감성 판단 임계값을 산출하는 작업이 필요하다. VOC의 경우 대부분이 불평과 불만에 대한 내용으로 부정적이었다. 각 문장별 가중치 합산 결과와 사람이 읽고 판단한 실제 감성을 이용하여 실제 감성과 비교하여 가장 높은 적중률을 보이는 임계값을 산출하였다. 그 결과 부정과 중립을 구분하는 임계값은 0, 중립과 긍정을 구분하는 임계값은 0.3으로 분석되었다.

6) VOC 데이터 감성분석 결과

이전 과정을 통해 구축한 사전과 사전 내 단어들의 가중치 그리고 두 가지 감성 판단 기준을 이용하여 검증 데이터 138개에 적용시켜 보았다. 두 가지 방법의 성능을 비교하기 위해 감성분석을 통해 판단한 감성과 실제로 사람이 읽고 판단한 감성이 얼마나 일치하는지를 비교하여 매칭률을 확인하였다. 감성분석 결과 아

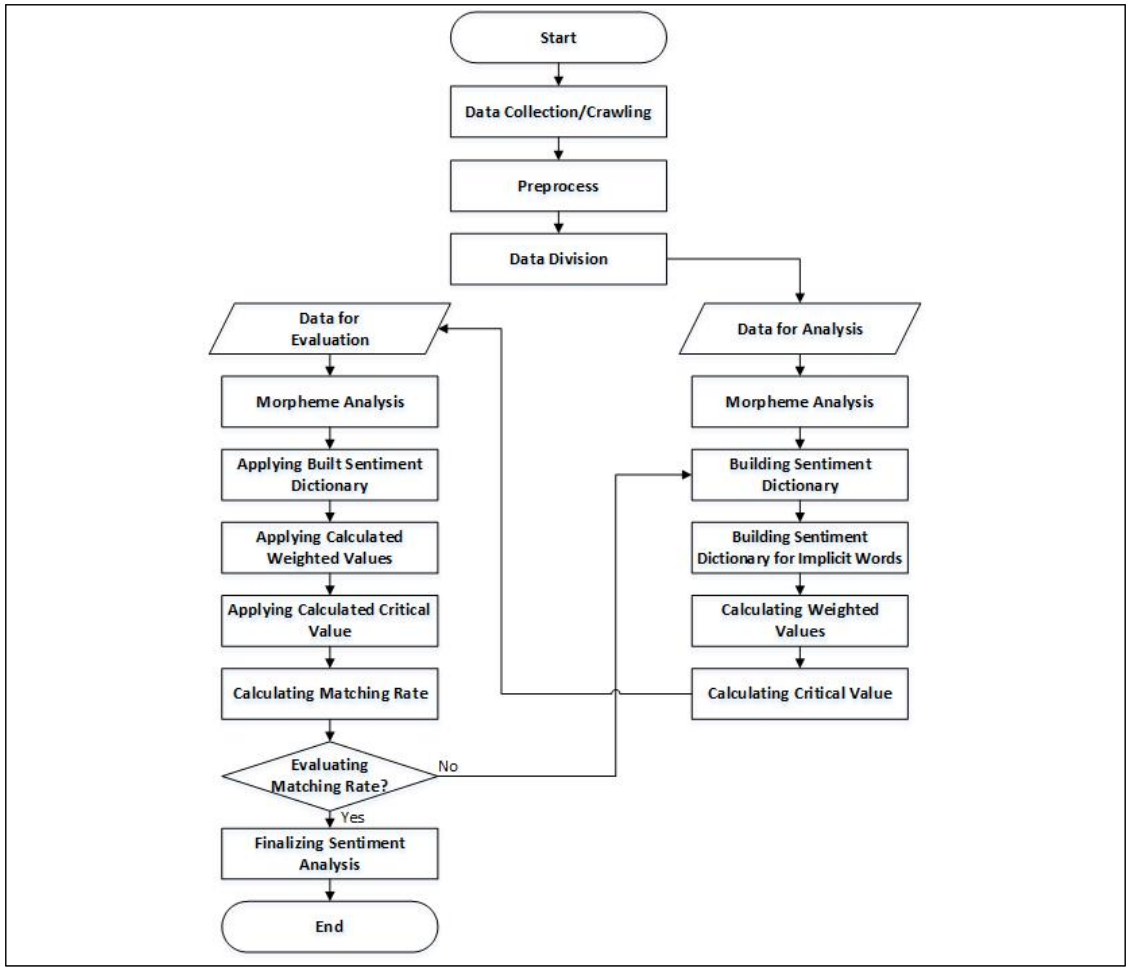
래 <Table 5>와 같이 주제별 의미를 이용한 방법이 범용적 의미를 사용했을 때 보다 실제 감성과 더 흡사하게 나타나는 것을 확인하였다. 이를 통해 주제별 사전을 구축하는 방법이 교통분야 감성분석에 더 적합한 것으로 판단되었다.

〈Table 5〉 Matching rate of Two Method

Classification	General Meaning	Subject Meaning
Matching rate	76.81%	80.44%

7) 교통정보 감성분석 방법론 수립

감성분석을 위해서는 자료 수집, 전처리, 감성사전 구축, 가중치 계산, 임계값 계산, 매칭률 계산, 감성분석 결과 평가 등의 과정을 거쳐야만 비교적 안정적이고 신뢰성 있는 감성분석 결과를 도출할 수 있다. 따라서 본 연구에서는 이렇게 수행한 과정을 바탕으로 <Fig. 5> 같은 감성분석 절차를 도출하였다.



〈Fig 5〉 Proposed Process for Sentiment Analysis

본 연구에서 제시하는 과정은 앞서 말한 감성 분석의 모든 과정을 처음 수행하는 경우를 고려하여 수립되었다. 즉, 신뢰성 높은 감성분석을 도출하기 위해 중요한 임계값, 매칭률 등에 대한 확인을 포함하고 있다. 추가적으로 수집된 자료를 이용하여 감성분석을 할 경우 <Fig. 5>의 검정용 자료에 적용한 방법과 같이 기존에 구축된 감성사전, 가중치, 임계값을 그대로 적용하여도 무방하다.

Ⅳ. 감성분석 방법의 적용성 검토

1. 고속도로 트윗 감성분석

1) 트윗 감성분석 개요

앞서 수행한 VOC 감성분석을 통해 교통 분야 감성분석에는 주제별 사전을 이용하는 것이 더 적절함을 확인하였다. 주제별 사전을 이용한 방법이 다른 데이터에도 적용 가능한지 확인하기 위해 SNS중 대표적인 트위터 데이터에도 적용하여 분석을 수행하였다. 트윗 감성분석을 위해서 “고속도로”를 키워드로 가진 트윗 9,590개를 수집하였다. 수집된 9,590개의 트윗에 0과 1사이의 난수를 임의로 부여하였고 0.9이상인 값을 가지는 트윗 943개를 검증데이터로 설정하였으며 나머지 8,647개의 트윗을 이용하여 주제별 사전을 구축하고 감성분석을 실시하였다(Hong, 2016).

2) 트윗 감성분석 결과

고속도로 관련 트윗으로 구축된 사전과 가중치를 이용하여 검증 데이터에 적용시킨 결과 고속도로 관련 트윗에 나타난 감성은 부정적인 것으로 분석되었다. 또한 분석된 결과에 대한 매칭률을 검토한 결과 <Table 6>과 같이 VOC 감성분석 결과와 비교하였을 때 다소 낮은 수치를 보였다. 그 이유는 대부분의 데이터가 단순 정보전달을 목적으로 하는 트위터에 비해 VOC의 경우 감성이 많이 드러나기 때문인 것으로 판단된다. 하지만, 트위터의 특성을 고려하였을 때 매칭률의 수준이 VOC 감성분석 결과와 비교하여 크게 떨어지지 않는 것으로 보아 방법론 정립에는 문제가 없는 것으로 판단하였다.

감성 사전을 구축한 결과 314개의 단어가 등재되었다. 사전 내의 단어들을 긍정과 부정으로 분류하였다. 그 결과 긍정 단어 103개, 부정 단어 211개로 분석되다. 또한 SNS의 특성 상 이모티콘의 사용빈도가 특히 높은 것을 확인할 수 있었다. 이모티콘을 제외한 감성 키워드 상위 10개를 분석한 결과 아래 <Table 7>과 같이 10개 중 8개의 키워드가 부정적인 표현임을 확인하였다. 이를 통해 고속도로 이용자들이 부정적인 감성을 많이 표출하고 있음을 확인할 수 있고, 고속도로에 대한 이용자들의 이슈가 지·정체, 사고와 같은 부정적인 상황에 집중되기 때문이라 판단된다. 이와 더불어 실제 고속도로 관련 트윗에 주를 이루는 내용을 알아보기 위해 형태소 분석 결과 중 일반 명사만을 이용하여 워드 클라우드(word cloud)를 구축하여 분석하였다. 그 결과는 <Fig. 6>과 같으며, 워드 클라우드를 통해 외곽순환고속도로, 서해안고속도로와 같이 장소를 표현하는 명사를 제외할 경우 정체, 차량증가, 소통원활, 공사 등의 고속도로 소통과 관련된 단어가 높은 빈도를 보였다. 이는 실제 고속도로 이용자들이 소통과 관련된 정보에 관심이 많은 것으로 판단된다(Hong, 2016).

〈Table 6〉 Matching rate of tweets

Classification	Subject Meaning
Matching rate	77.73%

〈Table 7〉 Top 10 sentiment keyword of tweets about Expressway

No	Sentiment	Keyword
1	positive	“괜찮”
2	negative	“막히”
3	positive	“한가”
4	negative	“위험”
5	negative	“시바”
6	negative	“밀리”
7	negative	“부글부글”
8	negative	“불편”
9	negative	“서운”
10	negative	“에이”

워드 클라우드 및 감성분석 결과를 통해 분석 기간 동안 고속도로와 관련하여 공사, 사고에 대한 정보 전달이 많이 이루어졌음을 짐작할 수 있으며 이로 인해 생긴 지체로 인해 이용자들의 불만이 높았던 것으로 판단된다.

〈Fig. 6〉 Word cloud of expressway related Tweets

2. 특정 장소 관련 트윗 감성분석

1) 트윗 감성분석 개요

앞서 분석한 트위터에서 제공되는 고속도로 교통정보에 대한 키워드 분석 및 감성 분석을 실시하였다. 그 결과 현재 제공되고 있는 주요 키워드와 일반이용자가 생각하는 특정 키워드에 대한 감정 여부를 확인하였다.

이러한 분석을 좀 더 세분화 하여 교통운영 및 전략도출에 활용하기 위한 분석을 실시하였다. 분석은 기존 수집된 트위터 자료에서 좀 더 세분화된 키워드이며 많이 언급된 서울외곽순환고속도로와 외곽순환고속도로를 이용하였다.

앞서 수집된 22,353개의 트위터 데이터 중 서울외곽순환고속도로와 외곽순환고속도로가 들어간 1,825개의 트윗 데이터를 이용하여 분석하였다. 1,825개의 트윗 중 공공기관이 작성한 트윗은 1,765개였으며, 일반인이

작성한 트윗은 60개로 나타났다. 분석방법은 본 연구에서 키워드 분석 시 구축한 키워드 사전과 감성 분석 시 구축한 감성사전을 사용하였다(Hong, 2016).

2) 트윗 감성분석 결과

서울외곽순환고속도로에 대하여 키워드 분석결과 “지체”가 545회, “정체”가 401회로 소통상황에 관련된 키워드의 빈도가 가장 높게 나타났으며 “사고”, “증가”, “소통원활”, “안전운행”, “공사”, “추돌사고” 순으로 나타났다.

이러한 분석은 단순히 트윗에서 지체, 혹은 정체의 단어가 몇 번이나 등장했는지에 대한 결과이므로 지정체, 정체, 지체와 같이 비슷한 의미를 전달하는 단어가 중복되어 표기된다는 단점이 존재한다. 따라서 이러한 비슷한 의미를 가진 단어를 하나로 표현했을 경우 어떻게 표현이 되는지 <Fig. 7> 및 <Fig. 8>과 같이 워드 클라우드를 통해 비교해 보았다(Hong, 2016).

〈Fig. 7〉 Keyword Analysis of the Seoul Expressway Highway (word frequency)

〈Fig. 8〉 Keyword Analysis of Circular Expressway in Seoul Outer (grouping similar words)

워드 클라우드 결과를 이용하여 서울외곽순환고속도로를 살펴보면 유사단어를 그룹화하였을 때 좀 더 많은 정보들이 나타나는 것으로 확인되었다. 이러한 부분은 앞서 감성분석 방법론에서 언급된 바와 같이 함축된 단어 및 유사단어를 고려하여 감성사전을 구축할 경우 보다 명확한 결과를 도출할 수 있음을 보여주고 있다.

Ⅴ. 비정형 교통 데이터 기반 SNS 모니터링 체계 활용 방안

1. SNS 모니터링 체계 활용 방안 개요

앞서 고속도로 관련 비정형 교통 데이터를 이용하여 감성분석을 수행하고 또한 분석 결과를 바탕으로 시사점을 도출하는 것이 의미 있음을 확인하였다. 이러한 결과를 바탕으로 본 연구에서는 비정형 교통 데이터를 이용한 교통 전반에 걸친 활용 전략으로서 교통 정보 전달 매체, 정보 수집 매체 그리고 정책 및 사용자 감성 모니터링 매체로서 활용하는 것을 제안하고자 한다.

1) SNS 기반 교통정보 제공 기능

SNS 기반 교통정보 제공 기능은 기존의 트위터 고속도로 교통정보 계정 등에서 수행하고 있는 기능을 말

한다. 하지만 현재 고속도로 교통센터와 같은 공식적인 교통정보 계정의 경우 생산하는 트윗의 수, 트윗의 다양성, 트윗 내용에 대한 만족도가 낮은 것으로 분석된다(Hong, 2016). 따라서 이 부분에 대한 기술적인 보완이 매우 필요하다. 이와 함께 비정형 교통 데이터 기반 모니터링 체계를 개발할 경우 상기와 같은 내용을 인력에 의존하지 않고 ITS 및 C-ITS와 같은 기존 시스템과 연계하여 자동적으로 수행할 수 있는 체계를 갖추는 것이 필요하다.

2) SNS 기반 교통정보 수집 기능

SNS 기반 교통정보 수집 기능은 기존에 없는 기능으로서 주기적으로 SNS 기반 비정형 데이터를 자동적으로 크롤링하고, 수집된 데이터에 대한 전처리 과정이 필요하다. 특히, SNS에서 생성되는 비정형 데이터들은 이모티콘 등 다양한 내용이 포함되어 있기 때문에 이러한 불필요한 정보들을 효과적으로 제고함으로써 필요한 정보의 손실을 막는 것이 매우 중요하다. 이를 통하여 특정 이벤트, 장소, 정책에 대한 실시간 모니터링이 가능할 것으로 기대된다.

3) 교통 관련 정책 및 이용자 만족도 모니터링

교통 관련 정책 및 이용자 만족도 모니터링 기능은 수집된 SNS 기반 비정형 데이터에 대하여 본 연구에서 제시한 방법을 이용한 키워드 분석, 감성 분석, 그리고 소셜 네트워크 분석을 자동으로 실행하고 그 결과를 리포트하는 것을 말한다.

키워드 분석 및 감성 분석을 위해서는 참조사전 및 감성사전을 잘 구비하는 것이 필요하다. 또한 구축된 사전들을 주기적으로 갱신하여 원하는 자료를 효과적으로 수집할 수 있도록 하는 것이 필요하다. 이때 함축적인 단어 및 유사단어에 대한 갱신도 함께 진행될 필요가 있다. 시대적 상황 및 계절 등에 따라서 SNS에서 언급되는 내용들이 달라질 수 있기 때문에 주기적으로 사전들을 재정비하는 것이 중요하다.

본 연구에서는 일상적인 모니터링을 염두에 두고 키워드 분석 결과에 대한 유형 구분을 “장소”, “교통상황”, “지시정보”로 구분하였다. 여기서 장소와 교통상황은 특정 이벤트가 발생하였을 때 기본적으로 필요한 정보이기 때문에 필요하다고 판단하였고, 지시정보의 경우에는 이벤트 발생에 따른 권고 행동을 제안할 필요가 있기 때문에 고려하였다. 하지만, 본 연구에서는 일부 기간 동안만의 트위터 데이터를 수집하여 분석하였기에 이루한 유형을 정확히 분류하기에는 한계가 존재한다. 따라서 SNS 모니터링 시스템이 구축되면, 지속적인 관찰을 통해 도출된 키워드를 유형별로 분류하는 작업을 시행해야 할 것으로 판단된다. 또한 일상적인 모니터링 외에 특정 정책이나 이벤트에 대한 심도 있는 모니터링의 경우에는 이벤트의 종류 및 분석 목적에 맞는 키워드의 유형분류가 필요하다.

Ⅵ. 결론 및 향후 연구

1. 결 론

본 연구에서는 SNS 중 하나인 트위터 감성분석을 통해 의미 있는 정보추출이 가능한지 확인하는 목적을 가지고 수행하였다. 이를 위해 한국도로공사 VOC 데이터를 이용하여 감성분석을 수행하였고 교통분야에 적합한 감성분석 방법론을 선정하였다. 그 결과 주제별 사전을 구축하는 방법이 교통분야에 더 적합한 방법임을 확인하였다. 선정된 방법을 토대로 트위터 감성분석을 진행하여 주제별 사전을 이용하는 것이 타당한 것으로 분석되었다. 또한 해당 트윗의 일반 명사를 이용하여 워드 클라우드를 생성하여 고속도로와 관련된 이

용자들이 의견을 파악할 수 있었다. 그 결과 고속도로와 관련된 의견 들 중에 부정적인 의견이 많이 표출되었음을 확인하였다. 그리고 분석 기간 내 고속도로의 보수공사, 추돌사고 등 소통에 영향을 주는 상황이 많이 발생하여 이로 인한 이용자의 불만이 많았음을 확인하였고 이용자들의 감성과 실제 사용된 단어들은 연관성이 있는 것으로 판단된다. 이는 곧 이용자들이 감성을 표출한 대상과 그 감성에 대해 분석한 결과를 통해 본 연구의 목적인 SNS 감성분석이 의미 있는 정보 추출이 가능한 것으로 판단된다.

마지막으로 본 연구에서 제안된 방법론을 바탕으로 비정형 교통 데이터를 이용한 교통 전반에 걸친 활용 전략으로서 교통 정보 전달 매체, 정보 수집 매체 그리고 정책 및 사용자 감성 모니터링 매체로서 활용하는 것을 제안하였다.

2. 향후 연구

본 연구에서는 교통분야에 적합한 감성분석 방법을 선정하였고 이를 트위터에 적용함으로서 의미 있는 정보 추출이 가능함을 확인하였다. 본 연구를 토대로 발전된 결과를 얻기 위해서는 다음과 같은 연구가 필요하다.

첫째, 좀 더 세분화된 주제별 감성 사전 구축에 관한 연구가 필요하다. 본 연구에서는 고속도로 전체에 대해서만 분석하였다. 향후 고속도로의 장소, 제공서비스 등과 같은 세부적인 주제별로 사전을 구축한다면 좀 더 나은 결과를 얻을 수 있을 것이다.

둘째, 카테고리별 주제어 사전에 대한 연구가 필요하다. 본 연구에서는 고속도로 전체에 대한 분석을 하였지만 고속도로 내에도 휴게소, 톨게이트, 소통 상황 등 세부적인 카테고리로 분류할 수 있다. 만약 세부적인 카테고리별 주제어에 대한 명사 사전이 구축된다면 해당 주제에 대한 감성분석과 더불어 특정 주제들에 대한 이용자들의 의견을 명확하게 파악할 수 있을 것이다.

셋째, 문장의 카테고리별 자동 분류에 관한 연구가 필요하다. 분석 대상이 되는 임의의 문장에 대해서 카테고리별 사전을 적용하기 위해서는 해당 문장 속에 숨어있는 정보를 통해 자동으로 그 문장이 어떤 카테고리에 속하는지 판단할 수 있는 기능이 요구된다. 향후 연구를 통해 위의 조건들이 모두 충족된다면 고속도로와 관련된 정책, 서비스 등에 대한 이용자들의 의견을 지속적이며 저비용으로 관찰할 수 있으며 요구사항에 대한 대책 수립에도 유용하게 사용할 수 있을 것이다.

넷째, 교통분야와 관련하여 트위터 등 SNS에서 빈번하게 사용되는 함축적인 단어(약어, 은어, 사투리)를 유형화하고 이러한 함축적인 단어에 대한 평가 방안을 모색할 필요가 있다. 이를 통해서 감성 분석을 이용한 정보 추출 방법론에 대한 표준화가 어느 정도 가능할 것으로 사료된다.

마지막으로 교통정책 시행 전후에 대한 평가 방법으로서의 감성분석 활용하는 것에 대한 연구를 추진할 필요가 있다. 본 연구에서는 단지, VOC 등을 이용한 감성분석을 사례로 들었으나, 교통 분야의 주요 사업 또는 정책에 대한 감성분석으로 분석분야를 확장시킬 수 있을 것으로 사료된다.

Text mining의 한 방법인 감성분석이 고객의 의견을 분석하는 데는 어느 정도 도움이 될 수 있지만, 해당 교통시설 이용자의 전체 의견을 대변할 수는 없다. 다만, 저렴하고 지속적인 모니터링 수단으로 활용도가 있을 것으로 판단되며, 앞서 언급한 위험성을 제거하기 위해서는 교통 분야에서 감성분석 적용에 대한 지속적인 연구가 필요할 것으로 판단된다.

ACKNOWLEDGEMENTS

본 연구는 한국연구재단 2010년도 및 2015년도 정부(교육과학기술부)의 재원으로 한국연구재단(NRF)의 지원을 받아 수행된 것임(NRF-2010-0029451, 2015R1A1A1A05028008).

본 논문은 홍두표의 박사학위논문에 게재되었던 내용을 수정·보완하여 작성하였습니다.

REFERENCES

Bae J. W., Son J. E. and Song M.(2013), "Analysis of Twitter for 2012 South Korea Presidential Election by Text Mining Techniques," *The Journal of Intelligence and Information Systems*, vol. 19, no. 3, pp.141-156.

Cho H. N., Chung Y. O. and Lee J. D.(2013), "Sentiment Analysis Using News Comments for Public Opinion Mining," *Proceedings of KIIS Spring Conference 2013*, vol. 23, no. 1, pp.149-150.

Godbole N., Srinivasaiah M. and Skiena S.(2007), "Large-Scale Sentiment Analysis for News and Blogs," *Proceedings of the International Conference on Weblogs and Social Media*, pp.219-222.

Google Our Planet 2013, Google, https://apac.th inkwithgoogle.com/intl/ko_ALL/, 2016.10.11.

Hong D. P.(2016), *Study on the Utilization of Unstructured Data for the Effective Management of Expressway Traffic Information*, Ajou University.

IDG Tech Report(2014), *Reading emotions in writing -Sentiment Analysis*, pp.1-10.

Kim Y. S. and Jeong S. R.(2013), "Intelligent VOC Analyzing System Using Opinion Mining," *The Journal of Intelligence and Information Systems*, vol. 19, no. 3, pp.113-125.

NH Economics Research Institute(2013), *Applica tions and Implications of Big Data*, pp.10-19.

Seniment Analysis, Naver, http://terms.naver.com/ entry.nhn?docId=2070770&cid=42346&categoryId=42346, 2016.10.11.

Seoul National University IDS Lab, http:// kkma. snu.ac.kr/, 2016.10.11.

Yu E. J., Kim Y. S., Kim N. G. and Jeong S. R.(2013), "Predicting the Direction of the Stock Index by Using a Domain-Specific Sentiment Dictionary," *The Journal of Intelligence and Information Systems*, vol. 19, no. 1, pp.95-110.

Yun H. M., Hong S. G. and Lee T. H.(2016), "Sentiment Analysis on Plant Business of Seawater Desalination in Gijang, Busan," *The Korean Association for Local Government Studies Conference Materials*, vol. 2015, no. 4, pp.13-20.

Zhuang L., Jing F. and Zhu X. Y.(2006), "Movie review mining and summarization," *Proceedings of the 15th ACM international conference on Information and knowledge management*, pp.43-50.