J. Soc. Korea Ind. Syst. Eng Vol. 41, No. 4 : 131-137, December 2018
https://doi.org/10.11627/jkise.2018.41.4.131
ISSN : 2005-0461(print)
ISSN : 2287-7975(online)

Unstructured Data Quantification Scheme Based on Text Mining for User Feedback Extraction

Jung-Heum Jo · Yong-Taek Chung · Seong-Wook Choi · Changsoo Ok†

Industrial Engineering, Hongik University

사용자 의견 추출을 위한 텍스트 마이닝 기반 비정형 데이터 정량화 방안

조중흠 · 정용택 · 최성욱 · 옥창수†

홍익대학교 산업공학과

People write reviews of numerous products or services on the Internet, in their blogs or community bulletin boards. These unstructured data contain important emotions and opinions about the author's product or service, which can provide important information for future product design or marketing. However, this text-based information cannot be evaluated quantitatively, and thus they are difficult to apply to mathematical models or optimization problems for product design and improvement. Therefore, this study proposes a method to quantitatively extract user's opinion or preference about a specific product or service by utilizing a lot of text-based information existing on the Internet or online. The extracted unstructured text information is decomposed into basic unit words, and positive rate is evaluated by using existing emotional dictionaries and additional lists proposed in this study. This can be a way to effectively utilize unstructured text data, which is being generated and stored in vast quantities, in product or service design. Finally, to verify the effectiveness of the proposed method, a case study was conducted using movie review data retrieved from a portal website. By comparing the positive rates calculated by the proposed framework with user ratings for movies, a guideline on text mining based evaluation of unstructured data is provided.

Keywords : Text Mining, Sentiment Analysis, Unstructured Data, Movie Review, Evaluation Framework

1. 서 론

기업 입장에서 자신의 제품 또는 서비스에 대한 고객의 평가는 향후 제품 또는 서비스를 새로 설계하거나 개선하는데 대단히 중요한 정보를 제공한다[5, 8]. 일반적으로 이러한 평가는 고객들에 대한 설문조사나 포커스 그룹 인터뷰 등을 통해 수집되어 사용자의 니즈 또는 제품의 요구사항을 도출하는데 활용되어 왔으나 비용이 많이 발생할 뿐만 아니라 분석에 명백한 한계를 가지고 있다. 설문조사의 경우 사전 준비된 한정된 질문에 대한 기준이 모호한 수치 정보만을 제공하고 포커스 그룹 인터뷰의 경우 평가자의 주관적인 견해가 반영되어 다소 편향된 결과를 도출하는 단점이 있다[6].

최근 인터넷 댓글 활성화 및 사회 관계망 서비스(SNS)의 확산으로 온라인에는 특정 제품, 장소 및 서비스에 대한 무수히 많은 개인의 의견 및 평가가 텍스트 형태로 존재한다. 인터넷 상에서 사람들은 수많은 제품, 사람, 서비스에 대한 많은 자신의 평가를 블로그, 커뮤니티 게시판,

Received 12 November 2018; Finally Revised 7 December 2018; Accepted 10 December 2018
† Corresponding Author : okcs@hongik.ac.kr

사회 관계망 서비스에 텍스트 형태로 작성하여 업로드 한다. 사실, 이러한 텍스트는 작성자의 제품 또는 서비스에 대한 중요한 평가를 담고 있어 경우에 따라서는 설문 조사 또는 인터뷰보다 통찰력 있고 중요한 정보를 제공할 수 있다. 그러나, 이러한 텍스트 기반 정보들은 비정형일 뿐만 아니라 정량적으로 수치화하기 어려워 제품 설계 및 개선을 위한 수학적 모형이나 최적화 문제에 적용이 어렵다는 한계를 가지고 있다. 다시 말해, 이러한 리뷰 또는 댓글 들은 작성자의 대략적인 생각이나 대상에 대한 대체로의 선호를 파악할 수는 있지만 그 제품을 얼마나 좋아하고 싫어하는지 정량적으로 평가할 수 없는 단점이 있다. 이와 같은 텍스트 기반 비정형 데이터를 정량화하는 방안으로 텍스트 마이닝을 고려할 수 있다. 텍스트 마이닝은 텍스트 형태의 정보로부터 작성자의 의견이나 포함된 정보를 추출하는 과정으로 단어 빈도수, 패턴 인식, 텍스트 클러스터링, 개념 추출 등의 기법으로 사람들이 작성한 글에 나타난 특정 주제나 대상에 대한 그 작성자의 주관적이고 감정적인 의견을 분석하는데 사용될 수 있다[1, 15].

따라서, 본 연구에서는 인터넷 또는 온라인에 존재하는 수많은 텍스트 기반 정보를 활용하여 특정 제품이나 서비스에 대한 사용자 의견이나 선호도를 정량적으로 추출하는 방안을 제안한다. 먼저, 특정 주제에 대한 사용자의 텍스트 기반 정보를 수집하고 이를 각 단어로 분해하고, 각 단어를 긍정, 부정으로 분류하여 최종적으로 해당 정보의 긍정률을 계산하여 사용자의 의견을 정량화하는 방안을 고려한다. 또한, 제안되는 방안의 적용 방안과 그 효과성을 입증하기 위하여 영화 리뷰와 영화 평점 평가를 활용한 사례 분석을 실시한다.

본 논문의 구성은 다음과 같다. 먼저, 제 2장에서는 텍스트 마이닝이 활용된 선행 연구에 대해서 설명하고, 제 3장에서는 텍스트 기반 정보의 정량화 방안을 제안한다. 그리고 제 4장, 제 5장은 제안한 방안을 영화리뷰에 적용한 사례 분석을 실시하고, 마지막으로 제 6장에서는 결론 및 향후 연구에 대하여 논한다.

2. 텍스트 마이닝을 활용한 비정형 데이터 분석

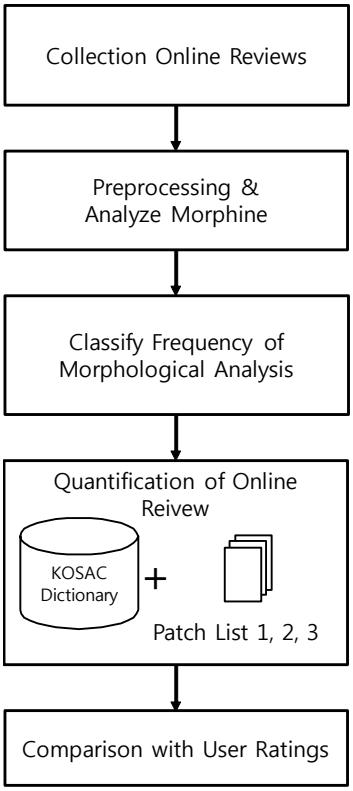
인터넷과 사회연결망 서비스(Social Network Service : SNS)의 발달로 방대한 양의 비정형 데이터들이 생산되고 있다. 비정형 데이터란 그림이나 영상, 음성, 문서처럼 구조화되지 않은 데이터로 최근 전 세계 기업에서 생성, 저장, 재사용하는 정보 중 80%는 복합문서(xls, ppt, doc, pdf)와 인터넷 페이지(html) 등의 비정형 데이터로 구성되어 있다고 알려져 있다[3]. 또한, IBM은 전 세계 데이터의 90%는 지난 2년 동안 생성되었고 신규 데이터의 80%는 비정형 데이터로서 정형 데이터의 2배에 달하는 속도로 증가하고 있다고 발표하였다[4]. 이와 같은 비정형 데이터 형태인 문서는 많은 유용한 정보를 포함하고 있으며 이 문서로부터 유효한 정보를 추출하고 가공하는 기술에 대한 요구가 점차 높아지고 있는 상황이다[3, 15].

텍스트 마이닝은 텍스트 형태로 이루어진 비정형 텍스트 데이터들을 자연어 처리 방식(Natural Language Processing)을 이용하여 가치와 의미가 있는 정보를 찾아내는 기술이라고 할 수 있다. 사용자는 텍스트 마이닝 기술을 통해 방대한 정보 뭉치에서 의미 있는 정보를 추출해 내고, 단어의 출현빈도, 단어 간 관계성 등 단순한 정보검색 그 이상의 결과를 얻어낼 수 있다[13]. 또한, 텍스트 마이닝 기법 중 하나인 감성분석(Sentimental Analysis)은 사용자가 작성한 문서에서 사용자의 감성과 관련된 텍스트 정보를 추출하여 문서를 작성한 사람이 어떠한 감성을 가지고 있는가를 판단하여 분석하는 기술로 비정형 데이터로 기술된 기사, 문서, 또는 자료 등에서 작성자들의 특정 주제나 제품에 대한 평가를 추출하는데 활용된다[9, 11].

Kam과 Song은 텍스트 마이닝을 활용하여 경향시문, 한겨레, 동아일보 세 개의 신문기사의 내용 및 논조 차이점을 단순 빈도 분석과 군집 분석, 분류 분석의 결과 비교를 통해 설명하였다[7].

제품 사용자의 주관적 의견을 자동으로 분류할 수 있는 감성분석 알고리즘은 상품에 대한 속성과 감성단어들에 대한 데이터베이스가 이미 구축되어 있다고 가정하고 이를 바탕으로 온라인 쇼핑몰에 등록된 한글 상품평에 대해서 전체 혹은 각 속성별로 긍정 또는 부정 의견인지 판단하였다[2]. 그러나, 이 연구는 상품평의 점수를 계산하기 위한 데이터베이스가 이미 구축되어 있는 것을 가정하고 있으며 온라인 리뷰에서 많이 사용되고 있는 인터넷 언어 혹은 신조어를 고려하지 않은 한계를 가지고 있다.

따라서, 본 연구는 SNS 텍스트를 기반으로 감성분석을 실시하여 비정형 데이터를 통하여 사용자 또는 생성자의 의견 또는 감성을 도출하는 방안을 개발한다. 인터넷으로부터 수집된 SNS 텍스트에 대하여 기존 단어 사전과 자체 개발한 형태소 사전을 활용하여 텍스트 마이닝을 분석하고 그 결과를 바탕으로 감성 분석을 실시하여 의미 있는 분석 결과를 제시하고자 한다. 특히, 자체 개발한 형태소 사전은 분석의 정확도를 높이기 위해 구어체 및 댓글체도 고려하여 설계되었다.



<Figure 1> An Evaluation Framework for Unstructured Data

3. 비정형 데이터 정량화 프레임 워크

본 장에서는 사용자가 작성한 비정형 데이터로부터 사용자의 감성 또는 평가를 수치화를 위해 <Figure 1>과 같은 텍스트 마이닝 기반의 비정형 데이터 정량화 프레임 워크를 제안한다.

제안되는 시스템은 먼저 비정형 데이터 수집하는 단계로 시작된다. 이 단계는 기존에 수집되어 있는 데이터를 활용할 수도 있고 크롤링(Crawling) 기법 등을 활용하여 인터넷에서 해당 데이터를 수집할 수도 있다. 특히, 인터넷에 많은 데이터 활용을 위한 크롤링 기법은 Web Scraping Technology를 이용하여 HTML 기반 웹사이트에 저장되어 있는 텍스트를 수집하는 과정으로 이를 반복, 수행하여 방대한 양의 비정형데이터를 획득할 수 있다.

비정형의 텍스트 정보가 수집되고 나면 이 후에 그 다음 단계로 수집한 텍스트에 대하여 결과에 영향을 주지 않는 불필요한 문자들(예) \/.,'"; 등의 기호들)을 제거하는 전처리 과정을 실행한다. 데이터 클린징이라고도 불리는 이 과정은 분석에 큰 영향을 주지 않는 조사 및 동사의 제거도 포함될 수 있다.

다음 과정으로는 한국어 정보처리를 위한 파이썬 패키지 코엔엘파이(KoNLPy)를 이용하여 수집된 정보를 형용사, 명사, 부사 등으로 분해하고 분해된 형태소별 단어를 긍정 또는 부정으로 분류한다. 이러한 분류를 위해 서울대학교언어학과 컴퓨터 언어학 연구실에서 개발한 KOSAC (Korean Sentiment Analysis Corpus) 감성 사전이 사용된다. KOSAC은 유일한 한국어 감성사전으로 총 1,600개의 주요 단어를 긍정, 부정으로 분류하고 있다. 앞서 분해된 단어를 이 단어 분류에 따라 긍정 또는 부정으로 분류하고 해당 문서 또는 데이터의 긍정 단어 수와 부정 단어 수를 계산한다. KOSAC에 포함된 1,600개의 단어는 댓글체, 급식체 등으로 대변되는 온라인 댓글 문화를 적절히 반영하기에 다소 제한적이므로 본 연구에서 KOSAC에 포함되지 않는 주요 단어를 대상으로 긍정 또는 부정으로 분류하는 3개의 추가 보완 목록(형용사, 명사, 댓글)을 구축하여 적용하였다. 마지막으로 주어진 문서 또는 비정형 텍스트 데이터에 대한 긍정률은 다음 식에 의해 계산될 수 있다.

$$\text{긍정률} = \frac{\text{긍정 단어수}}{\text{긍정 단어수+부정 단어수}} \quad (1)$$

앞서 설명은 전체 프로세스를 적용하여 주어진 문서에 대하여 이 긍정률을 계산되고 이는 해당 문서에 포함된 작성자의 감성 또는 의견이 판단하는 근거로 활용된다.

4. 사례 연구 : 영화 리뷰(댓글) 감성 분석

4.1 분석 개요

앞서 제안된 비정형 데이터 정량화 방안의 유효성을 확인하기 위하여 A포털사이트 영화 리뷰를 활용하여 사례연구를 실시한다. 특히, A포털사이트는 영화 리뷰와 함께 일반인(네티즌)들의 수치 기반 정보를 함께 제공있는데 이는 본 연구에서 제안된 비정형 데이터 정량화 방안의 결과의 정확도를 확인하는 데 활용될 수 있다. 분석을 위해 2017년부터 2018년 상반기까지 개봉하여 상영이 끝난 영화 중 30개의 영화를 무작위로 선정하여, 이 영화들에 대한 온라인 리뷰를 분석하였다.

먼저, 온라인 리뷰를 포털사이트로부터 네티즌들이 작성한 온라인 리뷰를 수집하기 위해 Python 기반 beautiful-soup과 request 라이브러리를 이용하여 크롤러(Crawler)를 개발하였다(<Figure 2> 참조). 이 크롤러를 활용하여 각 영화당 1,000개의 댓글과 평점을 수집하였고 댓글은 감성 분석을 위한 원 데이터로 평점은 산술 평균하여 감성 분석 결과와 비교하는 데 활용하였다.

```python
import urllib
import urllib.request
import urllib.parse
import bs4
import re
import os
import time
from concurrent.futures import ThreadPoolExecutor


def deleteTag(x):
    return re.sub("<[^>]*>", "", x)


def getComments(code):
    def makeArgs(code, page):
        params = {
            'code': code,
            'type': 'after',
            'isActualPointWriteExecute': 'false',
            'isMileageSubscriptionAlready': 'false',
            'isMileageSubscriptionReject': 'false',
            'page': page,
            'order':'newest'
        }
        return urllib.parse.urlencode(params)

    def innerHTML(s, sl=0):
        ret = ''
        for i in s.contents[sl:]:
            if i is str:
                ret += i.strip()
            else:
                ret += str(i)
        return ret

```

<Figure 2> A Python Program for Crawling

수집한 리뷰 분석에는 온라인 리뷰에서 흔히 나타나는 띄어쓰기 오류에 덜 민감한 한글 형태소 분석기인 꼬꼬마 형태소 분석기(Kkma, Kind Korean Morpheme Analyzer)을 사용하였다. 이 과정에서 문법적 오류가 심하거나 한글자 표현 등의 이유로 형태소 분석기가 분석하지 못하는 단어들은 제외되고 적합한 형태소를 추출한다. 이렇게 추출된 형태소 단위의 단어들은 KOSAC 감성 사전과 추가 보완 목록을 활용하여 긍정, 부정으로 분류되고 평가 대상인 비정형 데이터 또는 댓글에 대한 긍정 및 부정 단어 수를 계산한다. 이 단어 수들은 식 (1)에 의해 해당 댓글의 긍정률로 계산되고 이를 네티즌 평점과 비교하여 그 정확도를 측정한다.

4.2 감성 사전 보완 목록

앞서 언급한 바와 같이 KOSAC는 주요 1,600단어를 긍정, 부정으로 분류한 목록을 제공한다. 일부 단어의 경우 두 목록에 모두 포함되는 경우도 허용한다. 그러나 인터넷 댓글이나 리뷰 등으로 대표되는 비정형 텍스트 정보는 KOSAC이 분류한 1,600개보다 훨씬 많은 단어들이 포함된다. 따라서, 본 연구에서는 이러한 KOSAC의 한계를 극복하기 위하여 다음 3가지의 보완 목록을 제안한다. 첫 번째는 KOSAC에 포함되지 않은 형용사들에 대한 목록이다. 형용사가 텍스트 작성자의 감성을 잘 나타낼 것으로 예상됨으로 이에 대한 보완 목록을 고려한다. 두 번째 목록은 그 대상을 명사로까지 확대하여 형용사와 명사에 대한 보완 목록을 구성한다. 마지막은 사용된 형태소 분석기로 명사 또는 형용사로 분류되지 않은 신조어를 포함한 보완 목록을 구성한다. 이 목록은 Wikipedia에 명시된 '대한민국 인터넷 신조어'와 온라인 리뷰를 참조하여 구성하였다[14].

4.3 예제

KOSAC 감성 사전과 3개의 보완 목록이 어떻게 사용되는 지 다음 예제 댓글을 활용하여 설명한다.

"정말 재미없었습니다. 오늘 아침에 보고 왔는데 민망하기만 하네요 배우 분들은 좋으나 스토리 전개 코미디 모두 최하점 입니다. 올해 들어 최악의 영화네요. 너무 실망스럽네요."

먼저, 분해된 형태소를 KOSAC을 이용하여 긍정 및 부정으로 분류하면 다음과 같다.

- 긍정 단어 : {'배우', '하', '좋', '아침', '영화'}
- 부정 단어 : {'최악', '듣', '스토리', '보', '하'}

이에 따라 긍정률을 계산하면 5/(5+5) = 0.5가 된다. '-다'의 경우 생략되어 형태소로 분해되고, KOSAC만 이용할 경우 의미 있는 단어를 포함하지 않아 의미 없는 단어만 나타내는 경향이 있다.

두 번째로 형용사 보안 목록 리스트에 따른 분석 결과는 1개의 부정 형태소{'재미없'}가 추가로 검출되었다. 이를 바탕으로 긍정률을 계산하면 5/(5+6) = 0.455가 된다.

다음으로는 형용사와 명사를 추가한 목록을 사용할 경우 다음과 같이 더 많은 감정 관련 단어가 추출된다. 4개의 긍정 형태소{'코미디', '스토리', '정말', '모두')와 3개의 부정 형태소("최악", "실망", "민망")가 추가로 추출되었다. 이를 바탕으로, 긍정률을 계산하면 9/(9+9) = 0.5가 된다.

마지막으로 "형용사+명사+댓글체(신조어)" 목록의 경우 예시에서는 새로운 댓글체(신조어)가 사용되지 않아 긍정률은 이전과 같다. 하지만, 다음과 같은 단어를 추가하였다.

- 긍정 단어 : {'존잼', '쩜', '갓', '사스갓', '노빠꾸'}
- 부정 단어 : {'핵노잼', '개노잼', 안습', 롬곡옾눞'}

한글이 쓰이는 환경(인터넷)을 고려하여 영화 댓글의 생성과정 중 새로이 생기는 단어들을 감안하고 형용사, 명사, 댓글체(신조어) 이 세 가지에 포함되는 요소를 모두 추출하였다. 이와 같이 인터넷 언어를 이용한 온라인 리뷰에 대한 효과적인 분석을 위해 추가 보완 목록에 대한 연구 및 개발이 필요하다.

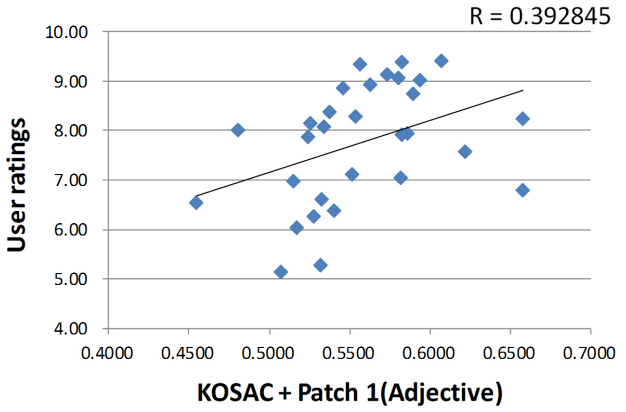
5. 실험 결과

제안된 비정형 데이터의 정량화 방안에 대한 검증을 위해 A포털사이트에서 2017년부터 2018년 상반기까지 개봉한 영화 중 30개를 임의로 선정하고 각 영화에 대한 댓글 1,000개씩을 수집하여 분석하였다. 수집된 댓글을 앞 장에서 설명한 3개의 보완 목록을 각각 KOSAC에 적용하여 분해하고 각각의 긍정률을 계산하였다. 각 목록의 유효성을 판단하기 위하여 각 영화에 대한 네티즌 평점과 상관분석을 실시하였다. 본 연구에서는 대체로 사용자들이 자신의 감성 및 의견에 따라 적절히 평점을 산정했다는 가정하에 검증을 실시하였고 추후 이 네티즌 평점의 정확성에 대한 별도의 논의가 필요하다고 판단한다.

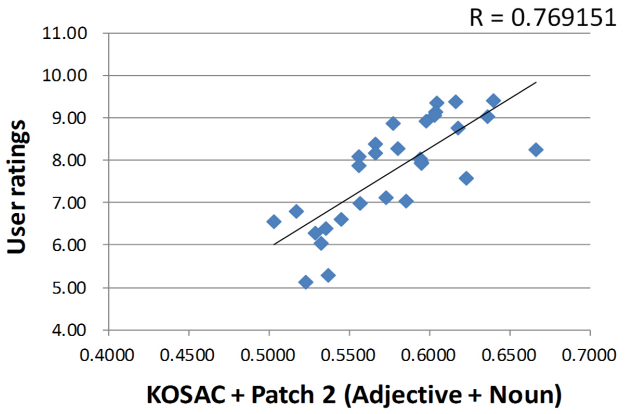
<Table 1>은 선정된 영화에 대한 3가지 평가 방법으로 계산된 점수와 네티즌 평점의 평균값을 보여준다. 높은 평점의 영화는 3가지 평가 방안으로도 높은 값을 가지고 반대의 경우도 마찬가지이다. 따라서, 본 연구에서 제안된 단어 기반 긍정률 계산이 유효한 것으로 판단될 수 있다. 더 자세히 분석을 위해 각 방안에 따른 긍정률과 사용자 평가 점수에 대한 상관 분석(Correlation Analysis)을 실시한 결과는 <Figure 3>, <Figure 4>, 그리고 <Figure 5>와 같다. 먼저, <Figure 3>에서 보는 바와 같이 사용자 평가 점수와 "KOSAC+형용사" 목록에 의한 점수의 상관 계수(Correlation coefficient)는 0.392845로 뚜렷한 양적 선형 관계를 가진다고 할 수 있다[10]. 두 번째로 "KOSAC+형용사/명사" 목록에 의한 점수와의 상관관계는 <Figure 4>에 나타내고 있으며 해당 상관계수는 0.769151로 강한 양적 선형 관계를 가진다고 할 수 있다[10]. 이는 영화 리뷰 텍스트 분석에서 형용사와 명사를 모두 고려한 평가 점수가 사용자 평점에 보다 유사하다고 할 수 있으며 사용자의 감성을 잘 반영한다고 할 수 있다. 마지막으로 "KOSAC+형용사/명사/신조어(댓글체)" 목록의 경우 가장 높은 0.797238의 상관계수값을 보이고 있으며 이는 텍스트를 분석할 때 특히 인터넷 텍스트 데이터를 분석할 경우 일상에 쓰이는 표준어 외에도 인터넷이라는 특수한 상황 속에서 쓰이는 단어를 고려해야 함을 시사한다.

<Table 1> Results with Three Patch Lists and User Ratings

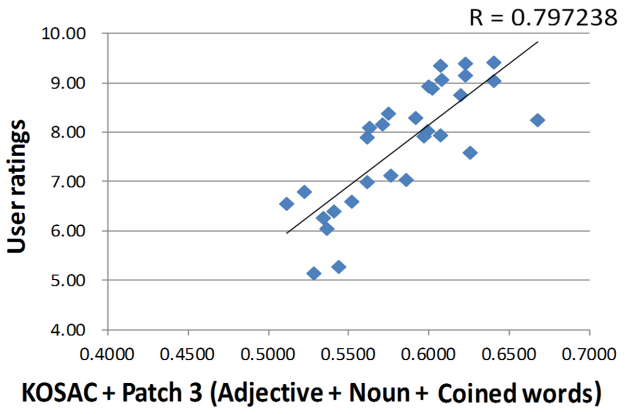
Movie	Patch 1	Patch 2	Patch 3	User Ratings
Wonder	0.6070	0.6392	0.6407	9.41
The Greatest Showman	0.5826	0.6162	0.6228	9.38
I Can Speak	0.5565	0.6040	0.6073	9.35
THE OUTLAWS	0.5733	0.6037	0.6226	9.14
A Taxi Driver	0.5804	0.6031	0.6079	9.06
⋮	⋮	⋮	⋮	⋮
Champion	0.5329	0.5449	0.5520	6.6
GONJIAM	0.4548	0.5029	0.5113	6.55
Asura	0.5280	0.5284	0.5344	6.27
The Battleship Island	0.5322	0.5370	0.5437	5.28
LOVE+SLING	0.5074	0.5228	0.5283	5.13



<Figure 3> User ratings versus KOSAC+Patch 1(Adjective)



<Figure 4> User ratings versus KOSAC+Patch 2(Adjective +Noun)



<Figure 5> User Ratings versus KOSAC+Patch 3(Adjective +Noun+Coined Words)

6. 결 론

사회 관계망 서비스, 블로그, 댓글 문화의 확산으로 인터넷에는 엄청난 양의 비정형 데이터가 생성, 저장되고 있다. 사실, 이러한 데이터는 사용자 또는 작성자의 의견이나 감성을 포함하고 있는 유용한 정보로 고려될 수 있으며 향후 제품 설계, 마케팅, 서비스 개발과 같은 주요 의사결정에 활용될 수 있다. 따라서, 본 연구에서는 이러한 비정형 데이터를 향후 의사 결정 문제에 활용하기 위하여 정량화하는 방안을 제안한다. 수집된 인터넷 문서 또는 댓글과 같은 텍스트 기반의 비정형 데이터를 형태소 분석기를 이용하여 최소 단위의 단어로 분해하고 이를 기존 감성 사전인 KOSAC을 이용하여 긍정률을 계산하여 얼마나 긍정적인 데이터인지를 평가한다. 이러한 긍정률 계산에서의 정확도를 향상시키기 위한 3개의 보완목록 (1) 형용사, (2) 형용사+명사, (3) 형용사+명사+댓글체(신조어)를 제안 적용하였다.

제안된 방안의 검증을 위해 임의 선정된 30개 영화에 대한 1,000개씩의 댓글을 수집, 분석하였고 그 결과를 포털사이트에 함께 제공된 평점과 비교를 하였다. 형용사로만 이루어진 첫 번째 목록의 경우 평점과 약 39%로 낮은 상관성이 나타낸 반면, 형용사+명사를 고려한 2번째 목록의 경우 76%의 상관성을, 품사의 형태가 다양하게 고려한 3번째 목록은 평점과 약 79%의 높은 상관성을 보여 댓글의 긍정도를 잘 평가하고 있은 것으로 판단할 수 있다.

이와 같이 본 연구에서 제안하는 비정형 텍스트 데이터 분석을 위한 정량화 방안은 온라인 댓글뿐만 아니라 블로그, 카페, 뉴스 기사, 기술 문서 등의 다양한 텍스트 데이터로부터 주요 의사결정에 필요한 정보를 추출하는데 효과적으로 사용될 수 있다. 본 연구는 최소 단위로 분해된 형태소가 갖은 의미를 바탕으로 전체 텍스트가 갖는 의미를 정량화하고 있다. 그러나, 일반적인 경우 텍스트의 경우 단어들의 순서나 조합이 갖는 별도의 의미가 존재하는 만큼 향후 이러한 부분을 고려한 감성 분석 방향에 대한 연구가 필요하다[12].

Acknowledgement

This study was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2018014267).

References

[1] Aggarwal, C.C. and Zhai, C.X., Mining Text Data, New York, Springer, 2012, pp. 11-35.

[2] Chang J., A Sentiment Analysis Algorithm for Automatic Product Reviews Classification in On-Line Shopping Mall, *Journal of Society for e-Business Studies*, 2009, Vol. 14, No. 4, pp. 19-33.

[3] Das, T.K. and Kumar, P.M., Big data analytics : A framework for unstructured data analysis, *International Journal of Engineering Technology*, 2013, Vol. 5, No. 1, pp. 153-156.

[4] Gantz, J. and Reinsel, D., The digital universe in 2020 : Big data, bigger digital shadows, and biggest growth in the far east, *IDC iView : IDC Anal. Future*, 2012, Vol. 2007, pp. 1-16.

[5] Ghose, A. and Ipeirotis, P.G., Estimating the Helpfulness and Economic Impact of Product Reviews : Mining Text and Reviewer Characteristics, *IEEE Transactions on Knowledge and Data Engineering*, 2011, Vol. 23, No. 10, pp. 1498-1512.

[6] Hu, M. and Liu, B., Mining and summarizing customer reviews, '*04 Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004, Washington, USA, pp. 168-177.

[7] Kam, M. and Song, M., A Study on Differences of Contents and Tones of Arguments among Newspapers Using Text Mining Analysis, *Journal of Intelligence and Information System*, 2012, Vol. 18, No. 3, pp. 53-77.

[8] Kim, K.A. and Ku, J.H., A Study on the Potential and Limitation of Pre-producing Dramas through Social Analysis-focusing on a jtbc drama <Man x Man>, *Journal of the Korea Academia-Industrial cooperation Society*, 2018, Vol. 19, No. 2, pp. 164-172.

[9] Kim, K.H., Chae, M.S., and Lee, B.T., Text Mining-Based Emerging Trend Analysis for e-Learning Contents Targeting for CEO, *Information Systems Review*, 2016, Vol. 19, pp. 2-4.

[10] Kim, S., Introduction to Statistics, Seoul, Hakjisa, 2007, pp. 96-97.

[11] Laudauer, T.K., Foltz, P.W., and Laham, D., An Introduction to Latent Semantic Analysis, *Journal Discourse Processes*, 1998, Vol. 25, No. 2-3, pp. 259-284.

[12] Le, Q.V. and Mikolov, T., Distributed Representations of Sentences and Documents, *Proceedings of the 31st International Conference on International Conference on Machine Learning*, Beijing China, 2014, Vol. 32, pp. 1188-1196.

[13] Tan, A., Text Mining : The state of the art and the

challenges, *In Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*, 1999, pp. 65-70.

[14] Wikidipia, https://ko.wikipedia.org/wiki/%EB% 8C%80%ED%95%9C%EB%AF%BC%EA%B5%AD%EC%9D%98_%EC%9D%B8%ED%84%B0%EB%84%B7_%EC%8B%A0%EC%A1%B0%EC%96%B4_%EB%AA%A9%EB%A1%9D(accessed on 11 November, 2018).

[15] Yoon, J., Song, J., and Ryu, T., Quantifying the Process of Patent Right Quality Evaluation : Combined Application of AHP, Text Mining and Regression Analysis, *Journal of Society of Korea Industrial and Systems Engineering*, 2015, Vol. 38, No. 2, pp. 17-30.

ORCID

Jung-Heum Jo | http://orcid.org/0000-0002-0555-4286
Yong-Taek Chung | http://orcid.org/0000-0002-9628-417X
Seong-Wook Choi | http://orcid.org/0000-0002-6752-3047
Changsoo Ok | http://orcid.org/0000-0002-2537-8160