소비자 분석을 위한 감성사전 모델링

이재웅*, 윤현노*, 문남미*
*호서대학교 컴퓨터소프트웨어학과
jaewoonglee@naver.com

Sentiment lexicon modeling for consumer analysis

Jae-Woong Lee*, Hyun-Noh Yun*, Nammee Moon*
*Dept of Computer Software, Hoseo University

요 약

본 논문은, 크롤링을 통해 얻은 비정형 데이터를 'Python'의 'KoNLPy' 라이브러리를 사용해 형태소 분석한 후 텍스트 마이닝을 통한 감성사전 구축을 목표로 하고 있으며, 형태소들의 빈도수를 기반으로 가중치로 두어 선별된 단어들을 이용해 긍정과 부정으로 나누어 카테고리화 한다. 이후, 선별한 카테고리에 단어의 극성을 판단하여 감성사전을 모델링한다. 실험을 위하여, 온라인 쇼핑몰 리뷰를 크롤링하여 비정형 데이터를 수집하고, 수집한 데이터를 분석, 가공 과정을 거쳐 정형화된 단어를 추출한다. 그 후에, 리뷰에 자주 사용되는 단어를 바탕으로 카테고리를 구성하였다. 구성된 카테고리 별로 단어의 극성을 판단하여 소비자 성향을 분석한 결과, 단순히 긍정과 부정을 표현하는 범용 감성사전보다 더 세분화된 감성 사전을 구축 할 수 있었다.

1. 서론

현대 사회에서는 스마트 디바이스가 발달함에 따라 온라인에서 소비자들은 물건을 구매하고 사용하며 그 정보를 공유하고자 상품 및 서비스에 대한 리뷰를 작성하고 잠재적 소비자와 많은 쇼핑몰 및 구매처에서는 상품에 대한 실 소비자의 정보를 얻고자 리뷰를 이용한다[1-2].

이처럼 해당 상품에 대한 정보는 기업 및 잠재적 소비자들에게 신뢰성 있게 받아들여지며 실용적으로 활용가능하지만 소비자가 쉽고 간단하게 리뷰를 작성할 수 있게 되면서 관련 데이터 또한 기하급수적으로 늘었고 데이터가 늘어남에 따라 자연어로 이루어진 비정형 데이터에서 필요한 정보를 얻기 어려워져 이를 해결하기 위한 방안으로 비정형 데이터를 정규화 된 데이터로 바꿔줄 필요성이 대두되었다[3-5].

대표적으로 비정규 데이터를 처리하기 위한 방법으론 텍스트 마이닝을 통한 감성 사전 및 오피니언 사전 구축 등이 연구되고 있으며[6-7] 나아가 두 사전들을 기반으로 한 소비자에게 상품을 추천하는 시스템 연구도 활발히 진행되고 있다[8-9].

기존 텍스트 마이닝 기법에는 형태소 분석 후 정규화 된 단어를 이용해 감성사전을 구축하였지만 범용적인 감성사전은 특정 분야에서 발생하는 감성단어에 대해서는 높은 감성분석 성능을 기대할 수 없기 때문에 양질의 감성사전을 구축하기 힘들었다. 본 논문에서는 이를 해결하고자 특정 분야를 미리 정했을 때를 가정하여 발생 하는 감성단어를 카테고리화 하여 기존 범용적인 감성사전에 비해 좀 더 정확하고 특정 요소에서 발생 할 수 있는 긍·부정 반응을 추출 가능하게 하는 감성사전을 제안한다.

2. 관련연구

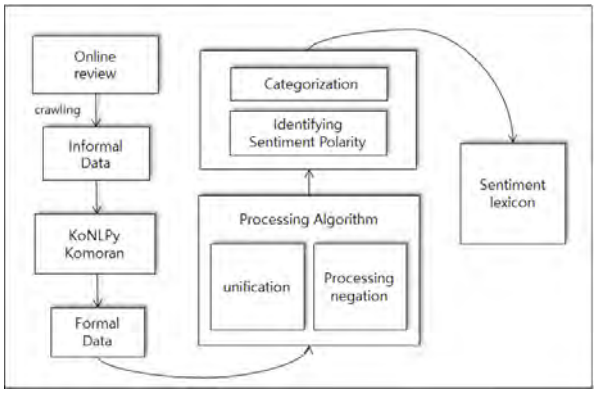
오피니언 마이닝이란, 웹사이트와 소셜미디어에 나타난 여론과 의견을 분석하고, 감성과 의견을 통계/수치화하여 객관적인 정보로 바꿀 수 있는 기술을 뜻하며, 이러한 분석을 이용하여 온라인 텍스트 속에 내포된 감성 및 감정을 식별하는데 유용하게 사용할 수 있다[6].

이러한 특징을 활용하여, 소비자를 위한 온라인 리뷰의 극성을 판별하거나, 상품 특징별 평가를 요약하는데 사용할 수 있으며, 이뿐 아니라 공급자에게 문제점과, 소비자의 불만 등을 분석하여 공급자에게 제공하는 방법에 대한 연구도 진행되고 있다[4].

또한, 비정형화된 텍스트를 정형화시키기 위하여 R 프로그램의 텍스트 마이닝 패키지(tm-Package)를 사용하여 온라인 리뷰를 분석한 연구가 진행되었다. 다만, 한글 단어와 패키지의 적합도가 떨어지고, 패키지의 정확도가 떨어지기 때문에 전체적인 결과의 정확도 역시 낮아지는 한계점을 보인다[8].

한편, 집단지성을 이용한 감성어 사전을 API와 결부시켜, 확장 활용하여 브랜드, 경쟁사, 평판 분석 등에 적용가능 할 수 있게 하고, 시간에 따른 변화를 감지 할 수 있게 하여 타 분야의 응용 방법을 제시하는 연구 또한 진행되었다[3].

3. 감성사전 모델링



(그림 1) 감성 사전 모델링 과정

(그림 1)은 본 논문에서 고안하는 감성사전 모델링 과정에 대한 순서를 보여준다. 먼저 Python 형태소 분석 라이브러리를 사용하여 형태소 분석을 실시해 비정형 데이터를 정형화 한다. 다음으로, 단어를 가공처리 한 뒤 카테고리화를 실시하고, 단어의 극성을 판별 하여 감성 사전을 구축하는 방식이다.

3-1. 형태소 분석

분석 자료를 수집하기 위해 데이터를 크롤링(crawling)한다. 이 때, 특정 상품에 대한 의견 및 평가를 수집하기 위해, 한 가지 대상에 대해 평가한 데이터를 모두 수집하여 한 개의 텍스트 파일 내에 전부 저장한다.

그 후에, 수집한 데이터를 'KoNLPy'를 사용하여, 비정형 데이터들을 단어 단위로 추출한다. 'KoNLPy'란 한국어 정보처리를 위한 Python의 오픈소스 라이브러리로, 'Twitter', 'Komoran', 'Kkma', 'Mecab' 등의 클래스를 제공하는데, 이 중에서 'Komoran'클래스는 총 42가지 tag로 데이터를 분류하여 적절히 세분화된 단어들을 제공하며, 'Kkma'처럼 더 많은 tag를 제공해주는 클래스에 비해 빠른 처리 속도를 보이기 때문에 본 연구와 적합하다[10].

3-2. 분석 단어 가공

<표 1> 감성 사전 형태소 분석 예시

품사	어미
'VV'(동사)	'~다'
'VA'(형용사)	
'NNG'(명사) + 'XS'(접미사)	
'NNG'(명사) + 'VC'(지정사)	
'XR'(어근)	'~하다'

다음으로, 분석한 단어에 대한 의미를 추출하기 위한 기준을 세운다. <표 1>은 추출된 단어를 사전에 정리하기 위한 기초적인 틀로서, 'VV', 'VA' 등의 태그를 분석기에서 수집한 뒤에 빈도수를 추출하기 위한 전처리 작업을 보여준다. '예쁜 것 같아요', '예쁘군요', '예쁩니다'처럼 같은 의미를 표현하지만 표현 방식이 조금 씩 상이한 여러 가지 단어에 대한 단일화를 위해 고안했다.

가공 과정에서 '부정어 처리'는 긍정과 부정을 판별해야하는 감성 사전에서 아주 중요한 부분을 차지한다. 예를 들어, '예쁘다'와 '예쁘지 않다'라는 두 가지 예는 완전히 상반되는 감정을 내포하고 있지만, 형태소 분석 결과 '예쁘(VV)'라는 동일한 결과로 추출된다. 이런 오류를 해결하기 위해 단순히 품사 뒤에 어미를 붙이는 과정 외에, '예쁘'라는 단어 앞뒤에 어떠한 단어가 나오는지에 대한 검사 과정이 필요하다. '~지않다'처럼 단어 뒤에 연결어미(EC) + 보조용언(VX)이 나오는 경우, '~이 아니다'처럼 부정 지정사(VCN)가 활용되는 경우, '안 ~하다', '별로 ~하다'처럼 단어 앞에 일반 부사(MAG)가 나오는 경우 등 여러 가지 경우를 고려한다.

3-3. 카테고리화 및 감성 처리

마지막으로, 분석한 형태소들의 빈도수를 기반으로 가중치를 두어 선별된 단어들을 이용해 긍정과 부정으로 나누어 카테고리화 한다. 이때, '슬프다'라는 단어는 감동적인 영화 리뷰에서는 긍정적 반응으로 볼 수 있지만 의류 관련 리뷰에서는 부정적 반응으로 볼 수 있기 때문에 분석하기 위한 데이터를 특정 분야로 한정시킬 필요가 있다. 이후 카테고리 내에 세부 카테고리를 구성하고, 각각 긍정 혹은 부정적인 의견에 해당하는 단어를 분류한다. 그 후 긍정적인 단어는 '+' 가중치, 부정적인 단어는 '-' 가중치를 두어, 단어의 극성을 판단할 수 있도록 한다.

4. 실험 및 결과

<표 2> 단어 빈도수 측정 결과

단어	개수	단어	개수
가격	2774	가볍다	1487
사이즈	2164	크다	1277
디자인	2078	주문	1252
예쁘다	1842	구입	1180
색상	1731	괜찮다	980
구매	1687	부드럽다	826
따뜻하다	1578	편하다	781
…	…	…	…

본 연구에서는 대표적인 온라인 쇼핑몰인 'CJ', 'GS', '현대' 쇼핑몰의 여성 아우터 리뷰를 바탕으로 감성사전을 구축하였다. 먼저, 리뷰 전체를 크롤링해 문서화 하여 총 10가지의 데이터를 추출했다. <표 2>는 추출한 데이터에서 카테고리를 만들기 위해 각 상품에 자주 등장하는 단어의 빈도수를 측정한 자료로, 개수 기준 상위 200개 단어로 데이터베이스를 구축했다. 자료를 바탕으로 소비자들의 의견을 분석하여 카테고리를 선별했다.

<표 3> 감성사전 모델링 결과

요소		극성	단어
가격		긍정	저렴하다, 싸다
		부정	비싸다
상품	디자인	긍정	예쁘다
		부정	별로다
	사이즈	긍정	맞다
		부정	크다, 작다
	…		
품질	소재	긍정	부드럽다,
		부정	거칠다
	마감	긍정	상태가 좋다
		부정	울다
	…		
…			

다음으로, 선별한 카테고리에 단어의 극성을 판단하여 사전을 구축했다. <표 3>은 구축된 사전의 일부를 보인다. 빈도수를 바탕으로 카테고리를 만들고, 세분화 할 필요가 있는 카테고리는 세부 항목을 만들어 해당하는 단어의 극성을 판별하여 분류했다. 이때 극성 판별은 가능하지만 카테고리로 분류되기 힘든 단어는 상품에 '기타' 항목을 두어 상품에 대한 전반적인 평가로 두었다.

<표 4> 구축된 감성 사전 바탕 리뷰 분석 결과

단어	카테고리	극성
"무겁다"	품질(무게)	부정(-1)
"따뜻하다"	상품(보온성)	긍정(+1)
"풍성하다"	상품(소재)	긍정(+1)
"좋다"	상품(기타)	긍정(+1)

<표 4>는 "좀 무겁긴 하지만 따뜻하고 풍성한 맛이 있어요. 좋습니다."라는 문장을 분석한 결과를 나타낸 표다. 위 문장을 분석한 결과, '무겁다', '따뜻하다', '풍성하다', '좋다'라는 단어로 추출됐다. 추출된 단어를 구축된 감성 사전에 대입하니 '무겁다-품질', '따뜻하다-상품', '풍성하다-상품', '좋다-상품' 카테고리로 분류됐다.

```
Identifying sentiment polarity review AND category

category_polarity[category1, category2, ...] = 0;
foreach word in review
    sentiment_lexicon(word);
    If(word == positive_word) then
        category_polarity[category] += 1;
    else if (word == negative_word) then
        category_polarity[category] -= 1;
end foreach

If (sum(all_category_polarity) > 0)
    review = positive review;
else If (sum(all_category_polarity) < 0)
    review = negative review;
else
    review = natural review;
```

(그림 2) 극성 판별 알고리즘

(그림 2)는 카테고리와 리뷰의 극성을 판단하는 알고리즘이다. 각 카테고리의 극성을 전부 0으로 초기화 한 다음, 분류된 단어들은 긍정일 경우 카테고리 극성에 +1, 부정일 경우 -1을 해주었다. 리뷰 전체의 극성은 모든 카테고리의 극성을 더한 값을 이용해 양수일 경우 긍정, 음수일 경우 부정, 0일 경우 중립으로 구분했다. 그 결과, 품질 항목은 -1, 상품 항목은 +3, 문맥상으로는 +2(-1(품목)+3(상품)) 이라는 결과가 나왔다. 이 리뷰를 남긴 소비자는 "상품에 관련해선 긍정적이었고, 품질에 관련해선 다소 부정적이었으며, 전체적으로, 긍정적인 리뷰를 작성했다."라는 결과를 도출할 수 있었다. 실험을 통해 한 리뷰에서 카테고리 항목별 극성과 문맥상의 극성을 종합적으로 판별할 수 있음을 확인했다.

5. 결론

지금까지 온라인 리뷰를 이용한 카테고리로 분류된 감성사전을 제안하였다. 감성사전 카테고리는 가공된 리뷰 데이터에서 얻을 수 있는 상품 관련 정보를 추출하여 구성하였다. 구성된 카테고리를 바탕으로 감성 단어를 분류해 감성사전을 구축하였다. 이를 통해 리뷰 내용의 극성을 판단할 수 있었으며 해당 상품의 요소마다 소비자의 긍·부정반응 또한 구분하여 얻을 수 있었다. 이는 기업에게 상품의 특성별 소비자의 반응을 파악할 수 있고, 잠재적 소비자에겐 자신과 비슷한 소비자의 추천 리뷰 서비스로 확장 가능하다.

이 감성사전은 특성상 제한된 분야에서 사용하는 목적으로 만들어 졌다. 때문에 원래 목적인 분야를 벗어났을 때 범용감성사전에 비해 성능이 떨어지는 경우가 발생한다. 이로써, 각기 다른 분야마다 맞는 감성사전을 따로 구축해야하는 한계점이 생긴다. 이를 해결하기 위해서는 범용감성사전과 카테고리로 분류된 감성사전이 적절히 조화된 감성사전 설계가 요구된다.

본 논문에서 제안하는 감성사전을 이용하여 소비자의 요구가 기존보다 세부적으로 반영된 추천 서비스 구현이 가능할 것으로 보인다. 나아가, 소비자가 과거 작성한 리뷰데이터를 바탕으로 소비자의 평가를 예측하는 서비스 등 확장된 연구에 대한 부분을 기대한다.

이 발표논문은 2017년 대한민국 교육부와 한국연구재단의 중견연구자지원사업의 지원을 받아 수행된 연구(한국연구재단-2017년-2017008886)임.

참고문헌

[1] 박은하, 김재규, 도주연, "사용후기가 온라인 제품 구매의도에 미치는 영향", 한국심리학회 학술대회 자료집 pp.582-583, 2007

[2] 이병철, 변효정, "온라인 리뷰가 관광상품 구매행동에

미치는 영향", 관광레저연구 26(7) pp.59-79, 2014
[3] 안정국, 김희웅, "한글 감성어 사전 API 구축 및 자연어 처리의 활용", 한국지능정보시스템학회 학술대회논문집 pp.177-182, 2014
[4] 연종흠, 이동주, 심준호, 이상구, "상품 리뷰 데이터와 감성 분석 처리 모델링", 한국전자거래학회 학술대회논문집, 2011
[5] 이종화, 이현규, "Data Dictionary 기반의 R Programming을 통한 비정형 Text Mining Algorithm 연구", 한국산업정보학회논문지 20(2) pp.113-124, 2015
[6] 김유영, 송민, "영화 리뷰 감성분억을 위한 텍스트 마이닝 기반 감성 분류기 구축", 한국지능정보시스템학회, 2016
[7] 정은희, 이병관, "오피니언 마이닝 기반 SNS 감성 정보 분석 전략 설계", 한국정보전자통신기술학회 논문지 8(6) pp.544-550, 2015
[8] 김주영, 김동수, "텍스트 마이닝 기반의 온라인 상품 리뷰 추출을 통한 목적별 맞춤화 정보 도출 방법론 연구", The Journal of Society for e-Business Studies pp.151-161, 2016
[9] 배경아, 인관호, 김응모, "감정분석을 통한 음악 추천 기법 연구", 한국지능정보시스템학회 학술대회논문집 pp.229-232, 2012
[10] 박은정, 조성준, "KoNLPy: 쉽고 간결한 한국어 정보처리 파이썬 패키지", 한글 및 한국어 정보처리 학술대회 논문집, 2014