감성분석 기반 호텔 리뷰의 특성별 극성 분석 및 유저의 선호도 반영 시스템

심하영*, 오수진**, 김응모*
*성균관대학교 소프트웨어대학
**성균관대학교 정보통신대학
e-mail : shyme2055@naver.com, bgbanana4@gmail.com, ukin@skku.edu

Aspect Based Sentiment Analysis System of Hotel Review, Reflecting User's Preference

Hayeong Shim*, Sujin OH**, Ung-Mo Kim*
*College of Software, Sungkyunkwan University
**College of Information and Communication Engineering, Sungkyunkwan University

요 약

인터넷을 통해 정보를 쉽게 공유하게 되면서 소비자는 제품이나 서비스를 이용하기 전 효율적인 의사 결정을 위해 먼저 작성된 다른 사람의 의견을 참고한다. 또한 기업은 이러한 소비자의 의견을 수집하여 제품의 피드백이나 마케팅 등 비즈니스적인 측면으로 활용한다. 수많은 상품평과 후기에서 특정 제품 또는 서비스에 대한 감성을 식별할 수 있다는 점에서, 감성분석은 소비자와 기업 모두에게 주목받고 있는 기술이다. 합리적인 결정을 위해, 소비자는 해당 웹사이트에서 제공하는 데이터를 참고하며, 이 데이터는 웹사이트마다의 기준에 따라 필터링된다. 하지만 제품/서비스에 따라 개인이 중시하는 부분이 다르기 때문에, 실질적으로는 다른 사용자의 의견을 참고하여 합리적인 결정을 내린다. 본 논문은 호텔의 리뷰를 여덟 가지 특성으로 구분하고, 각 특성별로 극성을 분석한다. 또한 사용자가 선호하는 특성에 가중치를 부여하여 순위를 나타내는 시스템을 제안한다. 극성분석 단계에서는 주어진 리뷰를 여덟 가지 특성으로 분류하고, 긍정/부정의 극성으로 분류하는 기계학습 알고리즘을 사용한다. 각각의 특성에 대해 가중치를 적용하여 얻을 수 있는 순서는 기존에 제공되는 순서보다 사용자의 선호도를 정확히 반영한다. 또한 본 논문의 제안을 호텔뿐만 아니라 다양한 제품/서비스에 적용하여 선호도를 반영한 순위 정보를 제공한다면 소비자의 합리적인 의사 결정에 도움을 줄 것이다.

1. 서론

1.1. 연구 배경

인터넷을 통해 정보를 쉽게 공유하게 되면서 소비자는 제품이나 서비스를 이용하기 전 효율적인 의사결정을 위해 먼저 작성된 다른 사람의 의견을 참고한다. 또한 기업은 소셜 네트워크 서비스에서 소비자의 의견을 수집하여 제품의 피드백이나 마케팅에 활용하는 등 비즈니스적인 측면으로 활용한다. 이러한 점에서 방대한 양의 데이터 속에서 소비자와 기업에게 유용한 정보를 이끌어낼 수 있는 감성분석이 주목받고 있다.

감성분석(Sentiment Analysis)은 제품, 서비스, 조직, 개인, 이슈 등을 다룬 텍스트에 내포된 의견, 감성, 평가, 태도 등을 분석할 수 있는 기술이다[1]. 상위 분류인 텍스트 마이닝(Text Mining)과의 가장 큰 차이는, 텍스트 마이닝은 텍스트에 내포된 '사실'에 초점을 두고 있으나, 감성분석은 텍스트에서 사용자가 취하는 '태도'에 집중한다는 점이다[2].

1.2. 연구의 필요성

많은 리뷰 사이트에서 사용자가 제공받는 제품/서비스의 순위 정보는 '별점순', '가격순' 등의 수치화된 일차원적인 데이터를 바탕으로 한다. 하지만 사용자의 결정에는 수치화된 데이터뿐만 아니라 개인의 취향도 관여한다. 예를 들어, 호텔을 예약하는 사용자는 객관적인 정보 이외에 위생 상태나 서비스, 편의 시설 등의 다양한 정보를 고려하여 실질적인 결정을 내린다. 감성분석을 기반으로 호텔 리뷰의 특성별 극성을 수치화하고, 사용자가 선호하는 특성에 가중치를 부여하여 순위를 나타낼 수 있다면 합리적인 의사 결정에 도움을 줄 것이다. 또한 이러한 서비스가 호텔 뿐만 아니라 다양한 제품/서비스에 적용 가능하며, 그로 인해 사용자가 얻을 수 있는 편리함이 크다고 예측하여 본 연구를 제안한다.

2. 관련 연구

2.1 감성분석 연구의 발전 배경

감성분석 연구는 컴퓨터 과학 분야의 자연어 처리 중 하나의 주제로서 연구되기 시작하여 현재는 다양한 학계 및 산업으로 확장되어 연구되고 있다. 이러한 확장은 제품/서비스에 대한 사람의 의견을 다양한 소셜 미디어를 통해 얻을 수 있고, 비즈니스적인 측면에서 대중의 의견을 분석함으로써 이윤을 극대화시키고자 하는 동기가 있기 때문에 발생한다[2].

[3]은 감성분석에 접근하는 다양한 연구 방법과 이론을 사실 기반의 분석에서 적용되는 다양한 방법론과 비교하며 정리한다. 감성분석은 본질적으로는 텍스트의 주제를 정치, 과학, 스포츠 등으로 분류하는 사실 기반의 분석에서 파생된 것이기 때문이다. 텍스트의 주제 분류는 주제를 나타내는 키워드에 집중하지만, 감성분석은 텍스트에 나타난 감성을 나타내는 단어에 주목한다는 점이 다르다.

2.2 기계학습 기반의 감성분석

감성분석은 크게 지도학습(supervised learning)과 비지도학습(unsupervised learning) 기반의 기계학습에 바탕을 두고 있다.

[2]에 따르면, 지도학습 기반의 학습 분류기 유형으로는 나이브 베이즈 분류기(Naive Bayes classifier), 지지벡터 분류기(Support Vector Machines), 결정트리 분류기(Decision Tree), kNN 분류기(k-Nearest Neighbors), 신경망 분류기(Neural Network), 최대 엔트로피 모델(Maximum Entropy) 등이 있다. 2002 년 Pang 과 Lee 에 의해 발표된 [4]는 지도학습 모델을 사용하여 영화 리뷰를 긍정과 부정의 두 가지로 분류한 첫 번째 연구이다. 이후 현재까지 다양한 연구에서 감성의 분류를 긍정/부정뿐만 아니라 긍정/부정/중립으로 나누고, 확률 언어 모델인 n-gram 에 변화를 주는 등 감성분석의 정확도를 높이기 위한 다양한 방법을 시도한다.

감성분석에서 지도학습 기반의 기계학습이 대다수를 차지하지만, [5]의 연구과 같이 비지도학습 기반으로 감성분석을 수행하는 경우도 있다. [5]에서는 의견을 나타내는 데에 빈번하게 쓰이는 구문의 패턴을 품사 태그의 집합으로 나타낸다. 그리고 연속된 단어의 품사가 이러한 구문의 패턴에 포함된다면 이를 추출하여 극성을 분석하는 방법을 사용한다. 구문의 패턴 중 한가지는, 형용사인 단어 뒤에 명사가 나오는 경우로 “This piano produces beautiful sounds.”에서 “beautiful sounds”와 같은 것이 포함된다. 그 외에 [6]과 같이 감성사전을 사용하는 방법도 비지도학습 기반의 감성분석에 포함된다.

최근에는 이러한 방법 외에도 딥 러닝(deep learning) 혹은 딥 뉴럴 네트워크(deep neural network)에 기반을 두는 감성분석 연구도 증가하고 있다.

2.3 특성별 감성분석

텍스트에 나타난 여러 가지 특성별로 감성분석을 수행하는 것을 Aspect Based Sentiment Analysis(ABSA)라 한다. [7]은 감성 분석에 관한 많은 연구가 해당 연구에서 다루는 도메인의 특징에 무관하게 일반적인 긍정/부정을 찾아내는 것을 문제로 인식한다. 그리고 음식점과 노트북이라는 주제를 설정하고, 주제별로 특징 단어를 파악하여 각각의 특징 단어에 대해 감성분석을 수행한다. 연구는 다음과 같은 네 가지 단계로 수행된다.

첫 번째는 특징 단어 추출(aspect term extraction) 단계로, 일련의 리뷰에서 다뤄지는 단어를 추출한다. 두 번째는 특징 단어 극성(aspect term polarity) 단계로, 특징 단어가 포함된 문장에서 긍정적인지, 부정적인지, 중립적인지를 판별한다. 세 번째 단계는 특성 카테고리 판별(aspect category detection) 단계로, 문장을 미리 정의된 일련의 카테고리(음식점의 경우 가격, 음식 등)로 나누는 단계이다. 네 번째 단계는 특성 카테고리 극성(aspect category polarity) 단계로, 문장이 포함된 카테고리가 긍정적인지, 부정적인지, 애매한지, 중립적인지를 판별한다.

본 연구에는 [7]에서 제시한 방법을 참고하여 호텔이라는 주제에 맞게 특성 카테고리를 설정하여 감성분석을 수행한다.

3. 데이터 수집 및 처리

3.1 데이터 수집

데이터 수집에 앞서, 본 논문의 비교, 분석이 될 웹페이지는 다음과 같은 기준에 따라 선정하였다. 1) 특정 제품/서비스에 대한 리뷰를 제공하고, 2) 관련 없는 내용을 최소화하기 위해 포스팅 형식보다는 댓글 형식의 리뷰를 제공하고, 3) 다양한 의견의 추출을 위해 리뷰의 개수가 많아야 한다.

트립 어드바이저에서 제공하는 리뷰의 길이가 적당하고 그 개수 또한 충분하다. 또한 상대적으로 타 사이트에 비해 인터넷 용어의 사용 빈도가 낮은 점에서 전처리 단계가 용이하다. 따라서 트립 어드바이저에서 서울에 위치한 호텔 100 곳의 후기를 최근에 작성된 순으로 50 개씩 수집한다.

SPA 로 구현된 웹페이지의 특성상 Chrome Web Driver 와 Selenium 을 사용하여 웹 크롤러를 구현한다. 이 과정에서 0.5 단위로 나누어져 있던 점수를 다시 계산하여 0.01 의 단위로 세분화하여 나타낸다.

3.2 전처리

전처리 과정은 Python 의 ‘re’ 라이브러리와 정규표현식을 사용하여 수행한다. 앞 단계에서 수집된 데이터는 크게 영어로 작성된 리뷰와 한국어로 작성된 리뷰가 있다. 본 논문에서는 영어로 작성된 텍스트 형식의 리뷰에 대하여 수치화를 진행하기 때문에 한국어로 작성된 리뷰는 영어로 번역될 필요가 있다. 한국어로 작성된 리뷰는 영어로 번역하기에 앞서 불필요한 한글 자음, 모음, 문장 부호 등을 제거한 후 구글에서 제공하는 번역 API 을 이용하여 영어로 번역한다. 영어로 작성된 리뷰는 번역 과정 없이 불필

요한 문장부호와 이모티콘 등을 제거한다.

3.3 수치화

수치화 단계는 전처리 과정을 마친 리뷰를 여덟 가지의 특성으로 나눠 분석한 후 각각의 긍정/부정 값을 수치화하는 단계이다. 여덟 가지 특성은 다음과 같다: 청결도, 편의 시설, 음식, 인터넷, 위치, 서비스, 가성비, 기타. MonkeyLearn 에서 제공하는 기계학습 기반의 API 인 호텔 카테고리 분류 API 와 긍정/부정의 극성을 분석하는 API 을 사용하여, 카테고리별 극성을 분석한다. 두 가지 API 에 대한 설명은 다음과 같다.

3.3.1 호텔 카테고리 분류기

호텔 카테고리 분류기는 먼저 호텔 예약 사이트인 북킹닷컴에서 뉴욕 호텔에 대한 리뷰를 수집한다. Amazon Mechanical Turk 을 사용하여 2000 개의 문장을 '청결도', '편의 시설', '위치'와 같은 일곱 가지 카테고리로 분류한다. 이를 SVM(Support Vector Machine)을 사용하여 기계학습 시킨 결과로 구현하여 호텔에 특화된 카테고리 분류기를 생성하였고, 약 80%의 정확도를 가진다.

3.3.2 호텔 리뷰 긍정/부정 분류기

호텔 리뷰에 대한 긍정/부정의 극성 분류기는 먼저 트립 어드바이저에서 뉴욕 호텔에 대한 리뷰와 별점을 수집한다. 별점이 3 초과인 후기는 긍정으로, 3 이하인 후기는 부정으로 라벨링한 데이터를 training samples 로 사용하여 기계 학습 모델링을 수행한다. 긍정과 부정 각각에 대해 5000 개 정도의 리뷰를 학습시키고, n-gram range 와 같은 변수를 조절하며 분석의 정확도를 높인다.

이러한 과정을 통해 만들어진 분류기는 호텔 리뷰에 특화된 분류기이다. 또한 카테고리와 긍정/부정을 분류하는 것에 그치지 않고 그 정확도를 0 과 1 사이의 값으로 제공하기 때문에 0 과 가까운 결과는 정확도가 낮거나 중립에 가까운 것으로 판단하여 분석 대상에 포함시키지 않을 수 있다. MonkeyLearn 에서 제공받은 API 을 Python 에서 다양한 케이스별로 사용한 결과, NLTK 와 SentiWordNet 을 사용하여 품사의 패턴을 바탕으로 극성을 분류하는 것보다 간단하고 정확함을 확인했다. 본 연구에서는 분류기를 사용하여 얻은 결과의 정확도가 0.6 이상일 때만 유의미한 것으로 분류한다.

4. 구현 및 결과 분석

4.1 구현

앞서 설명한 두 가지의 분류기를 사용하여 다음의 알고리즘으로 새로운 분류기를 구현하면 호텔에 특화된 ABSA 시스템을 얻을 수 있다.

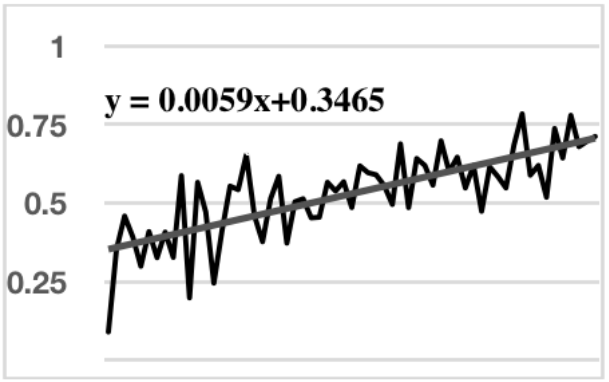
<표 1>에서의 3 번 라인과 같이 아무런 카테고리에 포함되지 않는 경우 array[7] 즉 '기타'에 긍정/부정의 극성값을 더한다. 5 번 라인과 같이 문장이 카테고리 하나에 포함되는 경우는 해당 array 에 긍정/부정의 극성값을 더한다. 7 번 라인과 같이 문장이 여러 개의 카테고리에 포함되는 경우는 'but', 'and', ','와 같은 단어로 나눈 결과에 대해 알고리즘을 수행한다.

<표 1> ABSA 시스템의 알고리즘

Algorithm ABSA	
	Input : A list of sentences *L*.
	Output : A polarity of the list *L*.
1	int[8] array ← {0,} *//polarity of each category*
2	**for** each line in L, **do**
3	**if** number of category == 0, **then**
4	array ← array + polarity *//polarity of 'General'*
5	**else if** number of category == 1, **then**
6	array ← array + polarity *//polarity of right category*
7	**else if** number of category > 2, **then**
8	**ABSA**(list of lines splitted by 'but', 'and', ',')
9	**return** sum of array[]

이를 통해 특정 호텔의 특성별 극성을 구할 수 있고, 극성을 합산한 순위와 트립 어드바이저에서 제공하는 별점 순서를 비교하여 새롭게 구현된 알고리즘의 정확성을 얻을 수 있다.

4.2 결과 분석



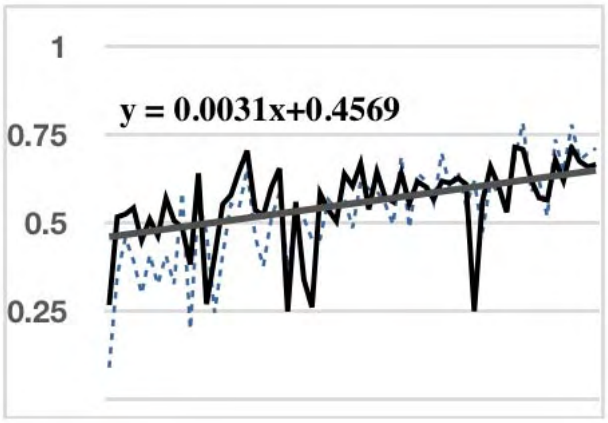
(그림 1) 별점순에 따른 극성 변화

(그림 1)의 그래프에서 x 축은 트립 어드바이저에서 제공한 별점 순서이고, y 축은 호텔의 특성별 극성값을 매긴 것이다. 전체적으로 별점이 높아짐에 따라 호텔의 특성별 극성값도 높아지는 것을 확인할 수 있다.

<표 2> 특성별 극성 비교

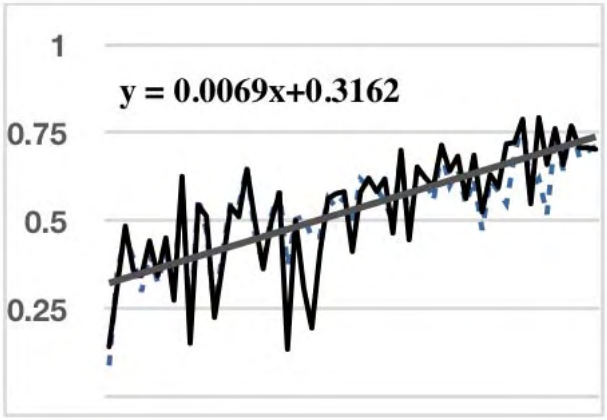
별점	위치	음식	청결도
2.71	0.70	0.64	-0.84
3.66	0.63	0.57	0.16
4.64	0.59	0.72	0.77

<표 2>는 별점이 각각 2.71, 3.66, 4.64 인 호텔의 위치, 음식, 청결도에 대한 극성값을 나타낸다. 부정에 대한 극성값에 음수를 취하여, 긍정과 부정의 극성값은 -1 과 1 사이의 값이다. 1 에 가까울수록 긍정, 0 에 가까울수록 중립, -1 에 가까울수록 부정적임을 나타낸다. <표 2>를 통해 별점이 높을수록 꼭 각 특성에 대한 극성값도 높아지는 것은 아니라는 것을 알 수 있다. 특히 위치에 대해서는 별점이 가장 낮은 호텔이 가장 높은 긍정의 극성을 보인다.



(그림 2) 위치에 가중치를 부여 했을 때의 극성 변화

(그림 2)의 그래프는 위치의 극성값에 가중치를 부여 했을 때를 나타낸다. 별점이 높을수록 위치에 대한 극성값도 높아지는 추세이다. (그림 1)과 같이 가중치를 부여하지 않았을 때의 추세선의 기울기와 비교해보면, 위치에 가중치를 부여했을 때 극성값의 증가폭이 상대적으로 작은 것을 확인할 수 있다.



(그림 3) 서비스에 가중치를 부여했을 때의 극성 변화

(그림 3)의 그래프는 서비스의 극성값에 가중치를 부여 했을 때를 나타낸다. 별점이 높을수록 서비스에 대한 극성값도 높아지는 추세이다. (그림 1)과 같이 가중치를 부여하지 않았을 때의 추세선의 기울기와 비교해보면, 서비스에 가중치를 부여했을 때 극성값의 증가 폭이 상대적으로 큰 것을 확인할 수 있다.

5. 결론

본 논문에서는 사용자가 선호하는 호텔의 특성에 가중치를 부여하여 순위를 제공하는 시스템이 호텔 예약 시 합리적인 의사 결정에 도움이 된다는 것을 제안한다. 웹 크롤러를 사용하여 트립 어드바이저에서 서울에 위치한 호텔 100 군데의 텍스트 형식의 리뷰를 수집하고, 전처리 과정을 거쳐, 지도학습을 기반으로 학습시킨 기계학습 분류기를 사용하여 호텔의 특성별 극성값을 구한다. 긍정에 대한 극성값이 높은 순서대로 순위를 매긴 결과 트립 어드바이저에서 제공한 호텔의 별점에 따른 순서와 비교하여 0.09 의 표준편차를 가진다. 또한 위치와 서비스, 청결도의 극성값에 가중치를 부여 했을 때는, 가중치를 부여하지 않았을 때와 비교하여 각각 0.11, 0.11, 0.16 의 표준편차를 가진다. 호텔의 특성에 가중치를 부여한 결과가 그렇지 않은 결과와 비교했을 때 차이를 나타내는 것이다. 이를 통해 사용자가 선호하는 호텔의 특성에 가중치를 부여하여 순위를 제공하는 시스템이 합리적인 의사 결정에 도움이 된다는 것을 확인할 수 있다. 차후 이를 호텔뿐만 아니라 다양한 제품과 음식점, 항공사 등의 서비스 분야로 확장시킬 수 있을 것이라는 점에 연구의 의의가 있다.

참고문헌

[1] Bing Liu, “Sentiment Analysis and Opinion Mining,” Morgan & Claypool Publishers, 2012.

[2] Appel, O., F. Chiclana and J. Carter, “Main concepts, state of the art and future research questions in sentiment analysis,” Acta Polytechnica Hungarica, Vol.12, 2015.

[3] Pang, B., L. Lee and S. Vaithyanathan, “Thumbs up?: sentiment classification using machine learning techniques,” Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, 2002.

[4] Pang, B., L. Lee and S. Vaithyanathan, “Thumbs up?: sentiment classification using machine learning techniques,” Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, 2002.

[5] Turney, Peter D., “Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews,” Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL-2002), 2002.

[6] Taboada, Maite, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede., “Lexicon-based methods for sentiment analysis,” Computational Linguistics, 2011.

[7] Liu, B., “Sentiment analysis and opinion mining,” Synthesis Lectures on Human Language Technologies, Vol.5, 2012.