

LDA를 이용한 온라인 리뷰의 다중 토픽별 감성분석* - TripAdvisor 사례를 중심으로 -

홍태호** · 니우한잉*** · 임강**** · 박지영*****

〈목 차〉

- | | |
|--------------------------------------|------------------------|
| I. 서론 | IV. 다중 토픽별 감성분석모형 및 실험 |
| II. 선행연구 | 4.1 데이터 수집 |
| 2.1 Latent Dirichlet Allocation(LDA) | 4.2 LDA기반 토픽추출 |
| 2.2 LDA기반 감성분석 | 4.3 제안모형의 토픽별 감성분석 |
| III. 연구 프레임워크 | 4.4 토픽별 감성분석 모형 검증 |
| | V. 연구결과 및 향후 연구과제 |
| | 참고문헌 |
| | <Abstract> |

I. 서론

소셜 네트워크의 급속한 성장으로 사람들은 여러 온라인 플랫폼에서 그들의 생각이나 느낌을 자유롭게 표현하고 공유하고 있다. 이렇게 사용자가 직접 생성한 콘텐츠(UGC, User-generated contents)는 폭발적으로 증가하고 있으며 여행지, 호텔, 식당 등과 관련된 사용자 리뷰도 이 중 하나이다(Ren and Hong, 2017). 인터넷 사용자들은 자신이 경험한 것들을 리뷰로

남기고, 또 다른 사용자들은 다른 사람이 작성한 콘텐츠를 통해 쉽고 편리하게 관심 대상에 대한 정보를 검색하거나 수집을 할 수 있게 되었다. 관련 조사를 살펴보면 여행과 관련된 정보 검색은 가장 인기 있는 온라인 활동 중 하나이며(Gretzel and Yoo, 2008), 사용자들이 남긴 온라인 리뷰에는 소비자의 체험, 사용자 피드백, 제품 특성 등의 내용이 포함되어 있기 때문에, 여행객들은 그들의 여행에 있어서 주 관심사인 여행과 관련된 온라인 리뷰를 적극적으로

* 이 논문은 2016년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임.

(NRF-2016S1A3A2925146)

** 부산대학교 경영학과, hongth@pusan.ac.kr(주저자)

*** 부산대학교 경영학과, niuhy90@gmail.com

**** 부산대학교 경영학과, mregan1314@pusan.ac.kr

***** 국민대학교 비즈니스 IT 전문대학원 BK21 플러스 사업팀, jiyoung.star@gmail.com(교신저자)

활용하고 있다(Litvin et al., 2008; Zhu and Zhang, 2010). 예를 들어, 여행자는 여행을 떠나기 전에 다양한 온라인 플랫폼(예, TripAdvisor.com)에서 다른 여행자가 남긴 의견과 경험을 확인하고 이를 참고해서 자기가 가고 싶은 호텔이나 레스토랑, 또는 관광지를 선택할 수 있다.

고객들이 표출하는 감정, 서비스에 대한 평가와 검토가 보다 명확해지고 있는 시점에서 비즈니스에 있어서도 고객리뷰에 대한 중요성은 더욱 커지고 있으며, 이를 활용하고자 하는 노력도 많아지고 있다(Ravi and Ravi, 2015). 즉, 그들의 제품과 서비스 품질을 개선하기 위해 고객들의 생각과 경험이 고스란히 담긴 온라인 리뷰를 분석하고, 그 결과를 활용하고자 하는 것이다(사공원 등, 2016; Duan et al., 2016). 온라인 고객리뷰는 여러 고객들에 의해 재생산되고, 재사용되면서 양은 더욱 방대해지고 내용은 보다 섬세해지고 있다. 그러므로 분석에도 많은 시간과 노력이 들기 때문에 결과를 적시에 활용하기가 어렵다. 이와 관련하여 최근 대량의 리뷰에서 고객들의 의견을 정확하게 추출하기 위해 자연어 처리 및 텍스트 마이닝과 같은 분석 기법의 개발에 관심이 집중되고 있다(김진화 등, 2011; Pang and Lee, 2008; Xianghua et al., 2013).

오피니언 마이닝이라고 불리기도 하는 감정 분석은 텍스트 마이닝의 한 부분이며 리뷰에서 유용한 정보를 추출하여 사용자의 성향이나 의견 등을 요약하고, 특히 리뷰의 주제에 대해 긍정적이거나 부정적인 표현들에 대해 분석할 수 있다(Su et al., 2008; Kim and Zhai, 2009). 감성분석은 문서수준(document-level), 문장수준

(sentence-level), 그리고 토픽수준(topic/aspect-level)의 분석으로 분류할 수 있는데(Liu, 2012), 문서수준의 감성분석은 문서 내 특정 용어들을 기반으로 감정을 표현하는 빈도수에 따라 긍정 또는 부정을 나타내는 문서를 분류하는 것이고, 문장수준의 감성분석은 문서에 포함된 문장이 긍정 또는 부정의 의견을 표현하고 있는지는 분석한다. 마지막으로 본 연구에서 집중하고 있는 토픽수준의 감성분석은 문서에 포함된 여러 토픽에 관련된 긍정 또는 부정의 극성을 탐색하는 것으로, 보다 세부적인 분석이라 할 수 있다(Hu and Liu, 2004; Lu et al., 2011). 현행 연구들은 주로 문서 및 문장 수준에서의 감성분석에 초점을 맞추고 있는데, 최근에는 토픽 기반의 감성분석 관련 연구가 증가하고 있는 추세이다(Lin and He, 2009; Lu et al., 2011; Marrese-Taylor et al., 2014; Nguyen and Shirai., 2015; Xianghua et al., 2013). 토픽 기반의 감성분석은 고객 리뷰에 대한 종합적인 평가 이외에 보다 세부적인 정보를 탐색할 수 있으므로 평가 대상에 대한 고객의 감성을 세부적으로 잘 파악할 수 있다는 이점을 제공하기 때문이다. 예를 들어 레스토랑에 관련된 리뷰는 일반적으로 음식, 환경, 분위기, 서비스 등과 같은 토픽에 대한 종합적인 평가로 볼 수 있다. 고객이 레스토랑에 대해 ‘전반적으로 좋았다’는 종합적인 평가를 했음에도, ‘음식 맛이나 분위기는 좋았으나 서비스는 별로였다’와 같이 각각의 토픽에 대한 감성은 다를 수 있다. 그러므로 고객들이 남긴 리뷰에 대한 심층적인 이해를 위해서는 종합적인 평가 이외에 각각의 토픽에 대한 기본적인 감성분석이 필요하며, 이는 고객의 의견이나 감정을 더 잘 파악할 수 있

는 방법이 되는 것이다.

본 연구의 목적은 온라인 리뷰에서 표출되는 고객들의 감성을 더욱 세부적으로 분석할 수 있도록 하는 온라인 리뷰에서 추출한 여러 토픽에 대한 감성분석 방법을 제안하는 것이며, 이를 위해 세계 주요 관광도시에 위치한 호텔의 리뷰 데이터를 수집하여 분석하였다. 이때 여행객의 리뷰에서 숨겨진 토픽을 추론하기 위해 Blei et al.(2003)가 제안한 Latent Dirichlet Allocation(LDA) 모델을 사용하였고, 추출된 각 토픽을 기반으로 감성분석을 실시하였다. 본 연구를 통해 여행객이 호텔에서 가장 중요하게 생각하는 요소는 무엇인지, 제안방법 및 결과물이 잠재고객 및 비즈니스 관리자들에게 어떻게 활용될 수 있는지에 대해 논의해보고자 한다.

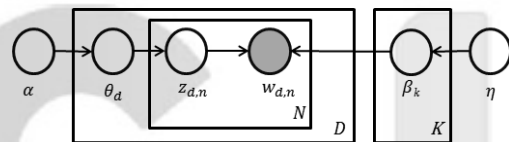
본 논문의 구성은 다음과 같다. 2장에서는 LDA 및 감성분석과 관련된 문헌을 소개하고 LDA 기반 감성분석 연구들의 연구방법을 설명한다. 3장에서는 본 연구의 제안모형인 고객이 직접 경험하고 작성한 호텔 리뷰의 토픽을 기반으로 한 감성분석 연구모형을 제시한다. 4장 실증분석에서는 수집한 데이터를 이용하여 토픽 기반 감성분석을 실시하고 결과를 분석한다. 마지막 5장에서는 연구의 시사점과 함께 연구의 한계 및 향후 연구문제에 대해 논의한다.

II. 선행연구

2.1 Latent Dirichlet Allocation(LDA)

LDA(Latent Dirichlet Allocation)는 비지도 학습 알고리즘으로 수많은 비구조적 문서에서

단어들 간 관련성에 따라 토픽별로 분류하는 확률적 토픽모델링 알고리즘이며(Blei et al. 2003), <그림 1>은 LDA의 문서생성과정을 도식화한 것이다. 색이 있는 노드는 관찰변수, 색이 없는 노드는 랜덤변수, 이외의 노드는 잠재변수를 나타내며, 노드 간 관계는 화살표를 통해 표현되고, 사각형은 각 문서, 토픽, 단어의 수를 의미한다. 먼저, 디리클레 분포를 따르는 토픽의 어휘 분포와 문서의 토픽 분포를 생성한 뒤, 각 문서 내 단어마다 토픽을 추출하고, 해당 토픽의 어휘 분포에서 단어를 추출하는 문서 생성 과정을 확률적으로 모델링 한 것이다.



여기서,

- K : 지정된 토픽의 수
- D : 문서의 수
- N : 문서의 길이, 단어의 수
- α : 토픽 K 의 Dirichlet prior weight, θ 값 결정 파라미터
- η : 토픽별 단어 w 의 Dirichlet prior weight, β 값 결정 파라미터
- θ_d : d 번째 문서의 토픽비율
- $z_{d,i}$: d 번째 문서에서 i 번째 단어의 토픽
- $w_{d,i}$: d 번째 문서에서 i 번째 단어(문서에서 추출되는 단어)
- β_k : 토픽 k 의 단어 w 의 생성확률
- k : 각 토픽 인덱스
- d : 각 문서 인덱스
- i : 문서 내 단어 인덱스

<그림 1> LDA 모형

여기서 LDA는 문서 D 와 각 문서의 단어 W 에 대해서 다음과 같은 생성 프로세스를 따른다.

다고 가정한다.

1. Choose $N \sim \text{Poisson}(\xi)$.
2. Choose $\theta \sim \text{Dir}(\alpha)$.
3. For each of the N words w_n :
 - (a) Choose a topic $z_n \sim \text{Multinomial}(\theta)$
 - (b) Choose a word w_n from $p(w_n|z_n, \beta)$, multinomial probability conditioned on the topic z_n .

LDA의 목표는 잠재 토픽 구조를 찾아내는 것이고, 관측값을 이용해 역으로 잠재된 토픽 구조를 추론하는 것이며, 관측변수들이 주어졌을 때 잠재변수들의 조건부 확률을 계산하여 이를 구할 수 있다.

실제로 문서를 작성할 때에는 이야기하고자 하는 주제를 중심으로 표현하게 되므로, 글 각 문서의 토픽 비율과 각 토픽 단어의 확률들을 알아낼 수 있다(Blei, 2012). 토픽모델링은 다양한 분야에서 활용되고 있는데 LDA기반의 토픽 모델링을 활용하여 문서 속에 표현된 문장, 단어들을 통해 전체 글에 대한 잠재적인 정보를 탐색할 수 있다. Song and Kim(2013)은 문장, 단어 등 관찰된 정보를 통해 문서에 숨어있는 정보를 추론하는 것을 목적으로 하여 생물정보학분야의 주제구조를 제시하였으며, 김상겸과 장성용(2016)은 LDA기반의 토픽모델링을 활용하여 산업공학에서 연구되는 주요 주제를 추출하고 변화를 분석하였다.

LDA 모형에서 전체 토픽의 개수 K 값을 결정해야 하는데, 이때 최적의 K 를 구하기 위한 방법으로 혼잡도(perplexity)를 들 수 있다. 혼잡도(perplexity)는 다음의 식(1)에 의해 계산되며(Blei et al., 2003; Zhao et al., 2015), 이 값이 낮을수록 LDA 모형이 텍스트를 분석할 때 우수한 일반화 능력을 나타낸다.

혼잡도:

$$\text{Perplexity}(D_{\text{test}}) = \exp \left\{ - \frac{\sum_{d=1}^M \log p(w_d)}{\sum_{d=1}^M N_d} \right\} \quad (1)$$

여기서, D_{test} 는 테스트 세트를 나타내며, M 은 D_{test} 가 포함된 문서의 개수, N_d 는 문서 d 의 길이, 그리고 $P(w_d)$ 는 모델에 의해 생성된 문서 d 의 확률을 나타낸다.

LDA를 사용하기 전에 먼저 결정해야 하는 토픽개수와 관련하여, Blei et al. (2003)은 적당한 토픽의 개수를 결정하기 위한 모델의 평가 지표로 표준 평가 기준 혼잡도(perplexity)를 선택하였고, Shi et al. (2015)은 LDA를 통해 분석을 할 때 토픽의 개수를 50개, 100개, 200개, 500개 4가지로 설정하고 키워드는 “top 5”를 선택하였다. 그리고 분류된 각 토픽 안에 포함된 키워드를 분석하여 토픽과 관련 있는 차원을 지정할 수 있다. 또한 LDA모델을 이용하여 토픽 개수를 선택할 때 K 값을 지정하는 데 있어 별다른 조건을 제시하지 않은 경우도 있다 (Gao et al., 2017). Cao et al. (2009)은 적합한 토픽개수 선택을 위해 토픽들 사이의 거리를 계산하여 토픽들의 관련성을 최소화하는 방법을 제안하였고, Zhao et al. (2015)와 Gao et

al.(2017)은 적당한 토픽을 개수를 선택하기 위해서 혼잡도(perplexity)와 교차 검증(cross-validation)을 사용하여 적합한 토픽을 결정하도록 하였다. 본 연구에서는 혼잡도(perplexity)와 교차 검증(cross-validation)을 결합하여 최적의 토픽 개수를 결정하였다.

2.2 LDA 기반 감성분석

감성분석(Sentiment analysis) 혹은 오피니언 마이닝(Opinion mining)은 자연어 처리(NLP, Natural Language Processing)의 중요한 작업 중 하나이며(Pang and Lee, 2008; Liu, 2012; Medhat et al., 2014), 온라인 상의 텍스트 데이터를 수집하고 분석하여 데이터 내에 잠재되어 있는 사용자의 감정, 태도 등을 유추하는 분석 방법이다(Liu, 2010; Yu, 2013, Wang, 2011). Hu and Liu (2004)는 감성분석을 사람들이 작성한 텍스트에서 감성을 분석해 평가 리뷰의 주제에 대한 감성이 긍정적인지 부정적인지에 대해 연구하는 분석기법이라 하였고, Liu (2012)는 감성분석이 오피니언 마이닝이며 텍스트 마이닝 기법의 일종으로 텍스트 속에 잠재되어 있는 감성을 탐색하여 중요한 정보를 얻을 수 있는 것으로 소개하였다.

감성분석 수준은 문서수준의 감성분석(Document-level sentiment analysis), 문장수준의 감성분석(Sentence-level sentiment analysis), 그리고 차원 또는 토픽수준의 감성분석(Aspect-level sentiment analysis) 등이 있다(Liu, 2015; Ren and Hong, 2017). 문서수준의 감성분석은 긍정적 또는 부정적인 의견 및 감성을 표현하는 문서를 분류하기 위해 문서 내 특정 용어들

을 기반으로 감성을 표현하는 빈도수가 판별 기준이 된다. 문장수준의 감성분석은 문서에 포함된 문장이 긍정적인지 부정적인지에 대해 분석한다. 마지막으로 토픽수준의 감성분석은 문서에 포함된 여러 가지 토픽과 관련된 감성을 분석하는 것이다. 이때 사전기반 접근법을 이용할 수 있는데 감성사전을 구성할 때 감성단어들의 감성수치를 미리 정의하여 연구 데이터에 있는 문서에 나타난 감성단어의 감성수치에 따라 리뷰의 감성값을 탐색하게 된다. Bravo-Marquez et al.(2014)은 수동으로 구축한 감성사전이 감성을 가지고 있는 단어들에 대한 문서의 극성 예측에 매우 유용하다는 실험 결과를 보였다. Mukherjee and Liu (2012)는 토픽수준 감성분석의 모델을 개발할 때 모형의 계수를 학습하기 위해 감성사전을 이용하였다(Hu and Liu, 2004).

특히 앞서 소개한 LDA 알고리즘을 기반으로 한 토픽 모델 관련 연구가 꾸준히 진행되고 있는 것을 볼 수 있는데, Lin and He(2009)는 LDA를 기반으로 한 새로운 확률적 모델링 프레임워크를 개발하였고, 이 연구 모델은 Joint Sentiment/Topic Model(JST)이라고 하며, 감성 분류와 토픽추출을 동시에 진행할 수 있는 토픽 모델이다. Xianghua et al.(2013)은 먼저 LDA를 적용하여 소셜 리뷰의 다중 글로벌 토픽을 추출하고 슬라이딩 윈도우의 컨텍스트(sliding window context)를 기반으로 지역적 토픽과 감성을 추출할 수 있음을 언급하였다. 이 연구에서 제시한 방법은 토픽분류와 감성분류에 대해 모두 좋은 성과를 나타내었다. Shi et al.(2015)은 LDA 모델을 활용하여 회사와 관련된 텍스트를 기반으로 토픽들을 추출하고 이에

대한 상위 5개의 단어를 기반으로 제품, 기술 및 생산, 마켓 등 세 가지 차원으로 분류하였다. 그리고 이 세 가지 차원의 분야에서 추가적인 방법을 통해 기업의 사회적 위치를 정량적으로 평가할 수 있음을 소개하였다. Marrese-Taylor et al. (2014)은 Liu(2007)의 방법을 개선하여 관광분야에 응용하였다. 이 연구는 차원 (aspect) 수준에서 더 복잡한 주관성분류와 감성분류의 자연어 언어 처리 규칙을 개발하였고 연구결과에서 제안 방법에 대한 효능성도 검증하였다. 이외에도 토픽수준의 감성분석(Topic-based sentiment analysis)은 여러 연구들을 통해 소개되고 있다. 문서의 문장을 기반으로 명사를 추출하여 토픽을 선정하고, 이 토픽에 따라 감성사전을 이용하여 감성분석을 하거나 (Nguyen and Shirai., 2015), 토픽의 개념을 토픽 수준으로 정의하고 각 문서의 차원을 분석하여 토픽을 추출하고 감성분석을 하기도 하는데, 이 과정에서 여러 전문가들이 상의를 거쳐 차원단어(aspect word)를 선정한다(Marrese-Taylor et al., 2014). 이외에도 토픽 모델과 감성분석을 결합하여 새로운 연구 방법을 제시하

는 연구도 있다(Lin and He, 2009; Lu et al., 2011; Xianghua et al., 2013).

<표 1>은 LDA 기반 감성분석 관련연구를 정리한 것이다. LDA 기반의 감성분석 관련연구를 보면 리뷰 데이터 내에 숨어있는 주제를 추출하고 동일주제에 대해 고객들의 심리동향을 탐색하려고 LDA 모델을 기본으로 새로운 알고리즘을 추가함으로써 연구를 진행한 것을 알 수 있다. 그리고 토픽 기반 감성분석 연구들은 전체 고객의 리뷰를 통합하여 추출한 토픽에 대해 전체적으로 어떤 감성인지 분석하는 연구들이다. 이와 같은 연구들에서 전체 리뷰 내의 잠재 토픽에 대해 판단한 감성 결과를 기반으로 전체적인 고객의 심리 동향은 표현할 수 있지만, 개별적인 고객의 심리 동향은 발견하기 어렵다. 하나의 리뷰가 반드시 하나의 감성을 표현한다고 볼 수 없으며, 하나의 리뷰 안에서도 여러 관점의 감성이 존재할 수 있으므로, 각 리뷰에서 토픽별로 감성분석을 한다면 사용자가 생각하는 것들을 조금 더 세부적으로 파악할 수 있게 된다. 고객들은 자신이 경험했던 상품에 대해 글로써 평가할 때 상품의 여러

<표 1> LDA 기반 감성분석 관련연구

연구	데이터	사이트	연구방법
Titov and McDonald (2008)	호텔 리뷰	TripAdvisor.com	Multi-Grain Latent Dirichlet Allocation model (MG-LDA)
Lin and He (2009)	영화 리뷰	cs.cornell.edu	Joint Sentiment/Topic model (JST)
Li et al. (2010)	상품 리뷰	Amazon.com	Dependency-Sentiment-LDA
Lu et al. (2011)	호텔 리뷰 레스토랑 리뷰	CitySearch.com OpenTable.com TripAdvisor.com	SVR, MG-LDA, STM
Xianghua et al. (2013)	중국 온라인 소셜 리뷰 (MSA-COSRs)	Sina Blog	Multi-aspect sentiment analysis (Sliding window 알고리즘)
Nguyen and Shiral (2015)	주식가격과 주식리뷰	Yahoo Finance	Topic Sentiment Latent Dirichlet Allocation (TSLDA)

측면에 대해 평론할 경우가 많은데, 이와 같은 경우 각각의 측면에 대해 고객의 감성을 탐색할 필요가 있는 것이다. 본 연구에서는 수많은 고객의 리뷰 내에 숨겨진 잠재적인 주제를 발견하기 위해 LDA 모델을 사용하였으며, 각 리뷰에서 토픽 별로 감성분석을 하는 모형을 제안하고자 한다.

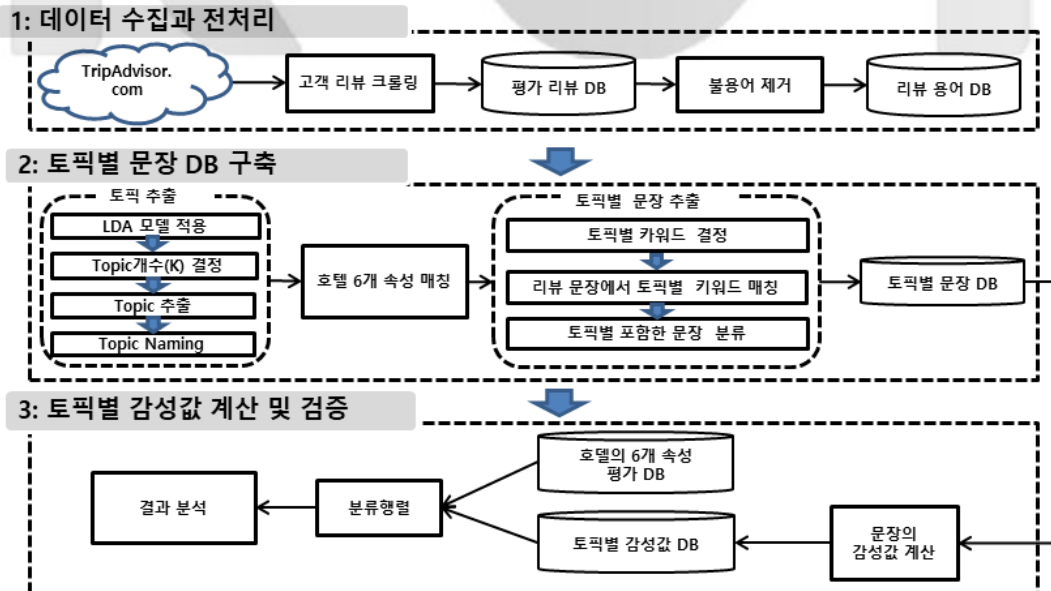
Ⅲ. 연구 프레임워크

본 연구의 프레임워크는 <그림 2>와 같으며, 데이터 수집과 전처리, 토픽별 문장 DB 구축, 그리고 토픽별 감성분류 및 검증 등 모두 3단계로 구성된다.

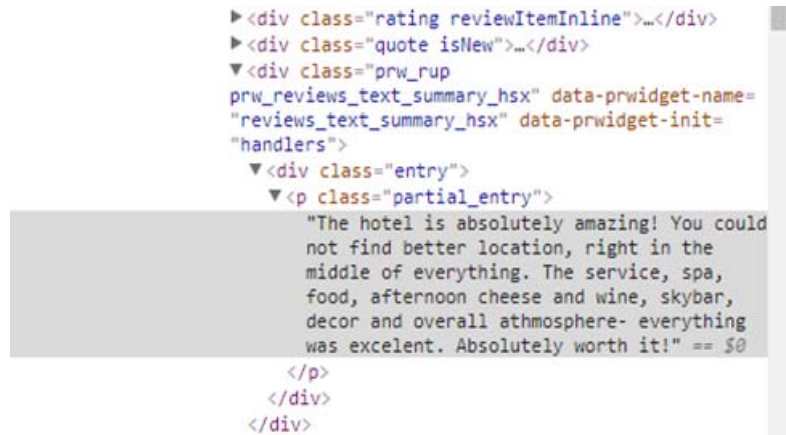
1단계는 데이터 수집과 전처리이다. 본 연구에서는 TripAdvisor 웹사이트에서 제공하고 있는 고객 평가 리뷰, 총 평가 점수와 속성별 평가

점수의 내용을 크롤링하고 정리해서 고객 평가 리뷰 데이터베이스를 구축한다. 수집된 데이터 내에는 결측치가 존재하고 리뷰 내에는 의미가 없는 불용어도 있을 수 있다. 2단계에서 보다 정확한 결과를 얻기 위해서 평가 리뷰 데이터베이스에 대해 전처리 절차를 거쳐야 한다. 전처리 이후 추출한 리뷰 용어 데이터베이스는 다음 단계의 실험에서 입력 데이터가 된다. 1단계 과정을 자세히 소개하면 아래와 같다.

실험에 사용할 데이터는 주로 웹사이트에서 사용자가 직접 생성한 콘텐츠(UGC, User Generated Content)를 크롤링하여 수집한다. 사용자 생성 콘텐츠를 수집할 때 필요한 데이터는 웹사이트의 페이지 소스(html)에서 찾아야 한다. 예를 들어 <그림 3>과 같이 고객 평가 리뷰가 저장되어 있는 페이지 소스의 위치를 찾아, 해당 고객의 평가 리뷰의 내용을 얻을 수 있다. 필요한 데이터의 페이지 소스 위치를 찾



<그림 2> 연구 프레임워크



```

<div class="rating reviewItemInline">...</div>
<div class="quote isNew">...</div>
<div class="prw_rup
prw_reviews_text_summary_hsx" data-prwidget-name=
"reviews_text_summary_hsx" data-prwidget-init=
"handlers">
  <div class="entry">
    <p class="partial_entry">
      "The hotel is absolutely amazing! You could
      not find better location, right in the
      middle of everything. The service, spa,
      food, afternoon cheese and wine, skybar,
      decor and overall athmosphere- everything
      was excelent. Absolutely worth it!" == $0
    </p>
  </div>
</div>

```

<그림 3> 고객 평가 리뷰의 페이지 소스 위치

아서 미리 설정된 R 프로그램 코드에 입력한 후, R 프로그램을 통해 새 코드를 이용하고 필요한 데이터를 자동으로 추출한다. 본 연구에서는 TripAdvisor.com에서 온라인 리뷰 데이터를 수집하여 실험에 사용하였는데, 수집 데이터는 세계 주요 관광도시의 호텔에서 숙박한 고객이 영어로 작성한 리뷰 데이터와 평가점수 데이터이다. 데이터를 수집한 후에 결측치가 있는 사례는 삭제하였는데, 고객이 웹사이트에서 평가를 하지 않은 부분이 있는 경우 해당 사용자 생성 콘텐츠는 이용하지 않았다. 고객 리뷰는 영어로 작성되었으며, 불용어를 제거하고 모두 소문자로 변환하여 처리하였다. 여기서 불용어란 리뷰에 포함되어 있는 특수한 문장 부호, 숫자, 웹사이트 주소가 빈번하게 나타나 의미가 없는 정보를 가리키는 용어이다.

2단계는 토픽별 문장 DB를 구축하는 과정으로 먼저 LDA를 이용하여 토픽을 분류하기 전에 최적의 토픽 개수를 결정해야 하는데, 혼잡도(perplexity)값과 교차검증(cross-validation)을 사용하여 최적의 토픽 개수를 결정할 수 있

다. 결정한 토픽 개수에 따라 수집된 리뷰 데이터를 대상으로 LDA 모델에 적용하여 중요한 토픽들을 추출한다. 추출된 키워드들의 전체적인 의미를 기반으로 분류된 토픽에 대해 해석한다. 그리고 추출된 토픽들을 TripAdvisor에서 운영하는 호텔들의 6가지 속성(value, cleanliness, rooms, service, location, sleep quality)과 대응시킨다. 대응되는 토픽의 키워드를 바탕으로 6가지 속성과 관련된 최종 키워드를 결정한 다음 토픽별 키워드를 포함한 문장을 대응시켜 토픽별 문장을 찾을 수 있다. 다음으로 각 리뷰의 토픽별 문장을 찾아서 토픽별 문장 DB를 구성하는데, 이 단계에서 토픽 모델 방법 즉, LDA 모델을 통해 토픽과 관련된 빈도가 높은 키워드를 자동으로 발견할 수 있다. 그리고 탐색한 토픽별 키워드가 존재하는 문장들을 토픽별 문장으로 분류한 후에 다음 단계인 토픽별 감성분석을 진행할 수 있도록 한다. 이 부분이 기존 연구와 본 연구 간의 큰 차이점이며, 본 연구 모형 중 새로운 절차에 해당된다. 기존 연구에서는 연구자가 미리 직접

토픽별 키워드를 결정하는 것을 볼 수 있는데, 이와 같은 방법보다 LDA를 통해 수집된 데이터 내에 출현 확률이 상대적으로 높은 단어들이 토픽별 키워드로 분류되는 것이 실질적으로 데이터의 정보를 더 잘 표현할 수 있다.

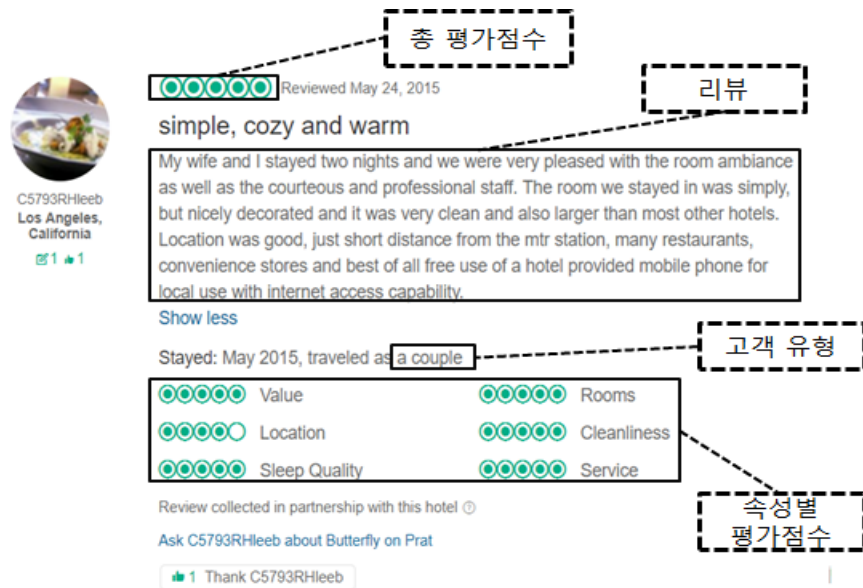
3단계는 토픽별 감성분석 결과 및 검증부분이다. 토픽별 문장 DB에 대해 감성 사전을 기반으로 감성분석을 진행해서 토픽별 감성분석 결과 DB를 얻는다. 관련 연구들을 살펴보면 토픽별 감성분석 방식이 다양한 것을 볼 수 있는데, 본 연구에서는 토픽별 키워드를 포함한 문장을 바탕으로 토픽별 감성분석을 진행하는 방식이 주요 프로세스이다. 감성분석을 할 때 고객의 감성결과에 대한 이진분류를 진행하고 분석한 감성값의 수치에 따라 긍정 또는 부정으로 분류한다. 온라인 리뷰에서 6가지 속성의 고객 평가내용과 제안 모형을 이용하여 계산된 각 리뷰의 토픽별 감성분석 결과를 비교하여 제안 모형의 성과를 검증할 수 있다. 추출된 토픽별 감성분석 결과가 유의한 값인지 아닌지를 검증하기 위해 분류행렬을 사용한다. 이때 정확도(accuracy), 정밀도(precision), 재현율(recall), F-1 measure 등 4가지 척도에 대한 결과를 보고 제안 모형의 성과를 평가하도록 한다.

IV. 다중 토픽별 감성분석모형 및 실험

4.1 데이터 수집

본 연구는 분석도구로 R언어를 사용하였는데, R의 'rvest' 패키지를 이용하여 TripAdvisor

웹사이트에서 세계 7대 도시(런던, 파리, 방콕, 베이징, 홍콩, 서울, 싱가포르)에 위치하는 호텔들의 온라인 리뷰를 수집하였다. 이 데이터는 2010년 1월부터 2017년 7월까지 고객들이 자신의 경험을 토대로 총 평가점수, 속성별 세부 평가점수와 평가리뷰를 작성한 것이다. <그림 4>는 TripAdvisor 웹사이트의 호텔에 관한 고객의 평가리뷰 내용이다. 본 연구를 위해 해당 사이트에서 총 평가점수, 리뷰, 고객 유형, 속성별 평가점수 등 4가지 내용을 크롤링하였다. 호텔과 관련된 다른 웹사이트와 달리 TripAdvisor 웹사이트에서는 value, cleanliness, rooms, service, location, sleep quality 등 6가지 속성별 평가점수를 찾을 수 있다. 본 연구모형의 성능을 평가하기 위해서는 속성별 평가점수 내용이 있는 데이터가 필요하기 때문에 TripAdvisor 웹 사이트의 고객 평가리뷰 내용을 이용하였다. TripAdvisor 웹사이트는 세계 최대의 여행 사이트이며 3억 1,500만 명 이상의 회원을 보유하고 있고 호텔, 레스토랑, 관광지 등 여행 관련 사업에 대해 5억 이상의 리뷰가 저장되어 있다. TripAdvisor 웹 사이트의 고객 평가점수는 최저값 1부터 최대값 5까지 다섯 개의 동그라미로 표시된다. 대부분의 리뷰 사이트에서는 각 리뷰의 평가 점수가 1부터 5까지 있는데, 평가점수 4와 5는 긍정적인 평가, 3은 중립, 그리고 1과 2는 부정적으로 분류된다(Fang and Zhan, 2015). 대부분의 연구 논문에서는 분류에 오류가 쉽게 생긴다는 이유로 중립 클래스를 사용하지 않는다(Liu, 2012). 본 연구에서도 온라인 리뷰의 점수 4와 5는 긍정적인 감성으로 간주하고 점수 1과 2는 부정적인 감성으로 지정한다.



<그림 4> 고객의 평가 데이터

<표 2> 수집된 데이터 요약(N=104,039)

rating	overall	rooms	service	value	location	sleep quality	cleanliness
1	1,866 (1.79%)	1,390 (1.34%)	1,012 (0.97%)	1,314 (1.26%)	444 (0.43%)	727 (0.70%)	888 (0.85%)
2	2,995 (2.88%)	2,235 (2.15%)	1,268 (1.22%)	2,110 (2.03%)	996 (0.96%)	1,031 (0.99%)	1,379 (1.33%)
3	10,777 (10.36%)	9,754 (9.38%)	4,623 (4.44%)	9,701 (9.32%)	5,429 (5.22%)	4,058 (3.90%)	5,202 (5.00%)
4	33,777 (32.47%)	23,697 (22.78%)	11,958 (11.49%)	24,895 (23.93%)	14,829 (14.25%)	11,152 (10.72%)	16,886 (16.23%)
5	54,624 (52.50%)	41,363 (39.76%)	26,451 (25.42%)	36,662 (35.24%)	32,519 (31.26%)	20,531 (19.73%)	43,198 (41.52%)
NA	0	25,600 (24.81%)	58,727 (56.45%)	29,357 (28.22%)	49,822 (47.89%)	66,540 (63.96%)	36,486 (35.07%)

수집된 호텔 데이터는 모두 104,039개 인데 그 중에 유효한 데이터는 모두 35,100개이다. 여기서 유효한 데이터란 6가지 속성에 대한 평가점수를 모두 포함하는 데이터를 지칭한다. <표 2>는 수집 데이터를 요약한 내용이다. 여기

서 “NA”는 고객이 해당 평가 항목에서 평가점수를 표기하지 않은 것이다. 6가지 속성에 대해 각각 평가점수를 표기하지 않은 비율이 상대적으로 높다는 것을 알 수 있으며, <표 3>에서 유효 데이터를 자세히 요약하였다.

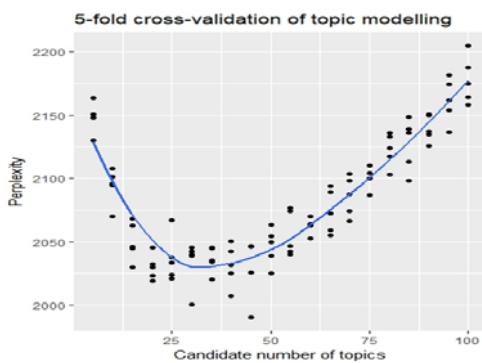
<표 3> 수집된 유효 데이터 요약(n=35,100)

구분	리뷰개수	비율	구분	리뷰개수	비율
호텔 종류 (hotel type)			도시(city)		
Budget	7,160	20.40%	Bangkok	5,088	14.50%
Business	13,622	38.81%	Beijing	4,430	12.62%
Luxury	14,318	40.79%	Hong Kong	5,939	16.92%
고객 유형 (traveler type)			London	5,693	16.22%
couple	14,404	41.04%	Paris	6,135	17.48%
business	6,200	17.66%	Seoul	2,838	8.09%
family	7,437	21.19%	Singapore	4,977	14.18%
friends	3,788	10.79%	계절(Season)		
solo	3,157	8.99%	spring	10,557	30.08%
no record	114	0.32%	winter	8,152	23.22%
			summer	7,681	21.88%
			autumn	8,710	24.81%

4.2 LDA기반 토픽추출

4.2.1 토픽개수의 선정

아래의 <그림 5>는 5-fold cross-validation을 결합한 토픽 수에 따른 혼잡도(perplexity)를 나타낸다. 그림의 기울기가 완만해지는 곳에서 토픽의 개수 30개를 지정할 수 있다. 설정된 토픽 개수의 범위는 2부터 100까지이다.



<그림 5> 토픽의 수에 따른 혼잡도(perplexity)의 비교

4.2.2 토픽 추출

수집된 데이터를 이용한 분석과정은 다음과 같다. 먼저 일반적인 리뷰 분석 절차에 따라 텍스트 자료에 대한 전처리를 거친다. 텍스트 전처리 과정에서는 분석에 필요하지 않은 불용어를 처리하고 숫자, 특수문자 등 분석하기에 곤란한 내용을 제거한 다음 LDA를 이용해 처리된 텍스트로부터 토픽별 키워드를 추출한다. 분석 프로그램으로는 R - 3.4.0(2017)을 사용하고 'tm,' 'lda' 등의 패키지를 이용하였다. LDA 모형을 이용하여 실험을 할 때 파라미터를 지정해야 하는데, α (alpha)값은 토픽 개수(K)를 50으로 나누어 얻은 결과로 지정할 수 있고, β (beta)는 0.01로 지정한다(Lin and He, 2009).

<표 4>는 30개 토픽들과 6가지 속성(value, cleanliness, rooms, service, location, sleep quality)의 매칭 결과이다. LDA모형에서 각 토픽에 대한 해석은 추출된 키워드를 바탕으로 판단하였는데, 예를 들어 토픽 1은 호텔의 프런트 데스크의 접대, 토픽 2는 룸과 욕실, 토픽 3

<표 4> LDA를 이용하여 토픽 추출

Topic ID	Dimension	Topic naming	Top 10 words
Topic 1	service	hotel reception	asked, told, reception, card, left, desk, called, call, booking, guests
Topic 2	rooms	bathroom and room	shower, bathroom, water, TV, bath, toilet, towels, bed, large, coffee
Topic 3	sleep quality	sleep environment	room, door, noise, air, open, floor, night, window, sleep, hear
Topic 4	service	breakfast	breakfast, coffee, food, buffet, fresh, tea, selection, fruit, restaurant, hot
Topic 5	service	dining	wonderful, beautiful, lovely, special, experience, amazing, service, dinner, suite, birthday
...
Topic 26	service	front desk	staff, service, front, desk, stay, made, concierge, make, booking
Topic 27	location	location	shopping, location, located, store, subway, front, food, convenient, restaurants, area
Topic 28	location	transport	taxi, train, river, road, back, night, station, taxis, skytrain, sky
Topic 29	service	room experience	service, business, quality, star, high, standard, experience, property, excellent, amenities
Topic 30	location	surrounding facility	walking, distance, restaurants, located, close, easy, location, area, shopping, city

은 고객의 수면 환경과 관련된 토픽이라는 것을 파악할 수 있다.

4.2.3 토픽별 키워드의 선정

호텔 리뷰에서 “staff”, “price”, “airport”, “downtown” 등의 단어들이 상대적으로 출현 빈도가 높다. “staff”은 호텔에서 제공한 “service”의 속성과 관련되어 있고 “airport”, “downtown”은 “location” 속성을 포함한 단어이다. 호텔의 평가 점수가 낮을 때는 일반적으로 고객의 평가 리뷰 중에 “bed”, “noise”와 같은 단어가 자주 나타난다. “bed”가 “room” 속성과 관련이 있으며, “noise”가 호텔의 “sleep quality” 속성에 영향을 미친다(Rhee and Yang, 2015). 평가 점수가 낮은 호텔뿐만 아니라 평가 점수가 높은 호텔이라도 “cleanliness” 속성을

중시해야 하며, 특히 이 속성은 관리자가 특별히 더 고려해야 하는 것이다(Stringam and Gerdes, 2010). 관광객들이 호텔을 선택할 때 최우선적으로 고려하는 속성은 “sleep quality” 속성인데, 호텔에서의 수면의 질에 대한 평가가 나쁘면 고객들은 절대 이 호텔에 묵지 않는다(Liu et al., 2013). 침대의 편안함, 조용함/방음 정도, 적절한 실내 온도 등이 “sleep quality” 특성으로 분류 될 수 있다. “location” 속성의 주요 요인은 “convenient”, “distance”, 그리고 “subway”, “bus”, “train” 등 교통수단과 거리 등이다(Rhee and Yang, 2015).

LDA를 통해 추출된 키워드를 바탕으로 6가지 속성과 관련된 최종 키워드를 찾았으며, 최종적으로 결정된 키워드 내용은 아래의 <표 5>와 같다.

<표 5> 각 속성 관련된 키워드

Dimension	Keywords
rooms	shower, bathroom, water, TV, bath, toilet, towels, bed, room, floor, facing, floors, beds, double, space, single, area, design, décor, style, suite, upgrade, wifi, internet, view, views, cleanliness, noise, air conditioning, smell, television, bug, carpet, minibar, refrigerator, amenity, furniture, wall, window, phone, light, kitchen
location	walk, station, minutes, road, minute, street, area, min, location, metro, tower, close, located, store, subway, convenient, airport, bus, mtr, mall, taxi, train, taxis, skytrain, sky, walking, distance, city, central
cleanliness	clean, dirty, dirt, smell, stain, broken, mold
value	free, charge, worth, price, cheap, budget, money, expensive
sleep quality	noise, night, sleep, quiet, soundproof, bed comfort, room temperature
service	staff, clean, breakfast, park, lounge, club, executive, drinks, wifi, internet, lobby, service, reception, guests, coffee, food, buffet, tea, fruit, restaurant, experience, dinner, front desk, concierge, amenities, pool, orchard, gym

4.3 제안모형의 토픽별 감성분석

4.3.1 감성분석

리뷰의 감성분석은 기본적으로 문서가 긍정, 부정, 또는 중립 중 어떤 견해를 갖고 있는지 판별하는 과정이다. 감성분석은 감성사전을 이용하여 진행되고, 각 문서 최소 단위인 어휘의 감성 극성(Sentiment Polarity)에 기반하여 이루어진다. 본 연구는 Bing Liu 사전을 사용하여 호텔 리뷰에 대해 감성분석을 실시하였다. R 프로그램의 ‘syuzhet’과 ‘dplyr’ 패키지를 이용하여 토픽별로 문장들에 대한 감성분석을 실시하였는데, 감성 분석 후 감성 값에 따라 문서의 감성분류에 차이가 생긴다. 예를 들어 감성분석에서 0값은 고려하지 않고, 0값보다 크면 긍정적으로 분류하고 작으면 부정적으로 분류한다(Hu and Liu, 2004; Keshavarz and Abadeh, 2017). Bravo-Marquez et al.(2014)의 경우에는 각 사전을 이용해 분석을 할 때 0값을 중립적인 감성으로 지정하였다. 본 연구에서는 감성값 0은 고려하지 않고 고객의 긍정(감성값>0)과 부

정(감성값<0)에 대해서만 분석한다. 토픽별 감성분석 과정은 아래 <그림 6>과 같다.

```

sentiment(reviews D, lexicon B)
For each review  $d_i$  in D
  score = 0
  For each  $w_j$  in  $d_i$ 
    score = score +  $v_B(w_j)$  //
    the score of  $w_j$  in lexicon B
  end for
  if (score > 0)
    sentiment = positive
  else if (score < 0)
    sentiment = negative
Return sentiment

```

<그림 6> 감성분석 알고리즘

<그림 6>에서 lexicon B는 감성사전이며 리뷰 데이터는 reviews D로 표시한다. W_j 는 리뷰 안에 포함된 감성 어휘이다. 각각 리뷰의 score가 0보다 크면 긍정적으로, score가 0보다 작으면 부정적으로 분류한다.

감성사전에서 각 단어의 감성수치를 미리 정하는데 부정적인 단어의 수치는 음수로, 긍정적인 단어의 수치는 양수로 설정한다. 본 연구에서 사용된 감성분석 알고리즘은 다음의 2단계

로 해석될 수 있다. 먼저 문서가 포함하고 있는 여러 가지 부정단어의 수치와 긍정단어의 수치의 합을 계산한 다음 이 값을 0과 비교하는데, 0보다 크면 긍정적인 감성으로 그리고 0보다 작으면 부정적인 감성으로 분류한다. 본 연구에서는 사용자 리뷰 35,100개를 대상으로 추출한 모든 긍정/부정 단어들에 대하여 감성분석을 진행하였다.

감성분석의 예시를 들면 아래 과정과 같다. 리뷰의 감성값은 총감성값이 -5로, 부정으로 판단하고, 이러한 과정을 총 35,100개의 리뷰에 적용하여 리뷰별 감성분류를 수행하였다.

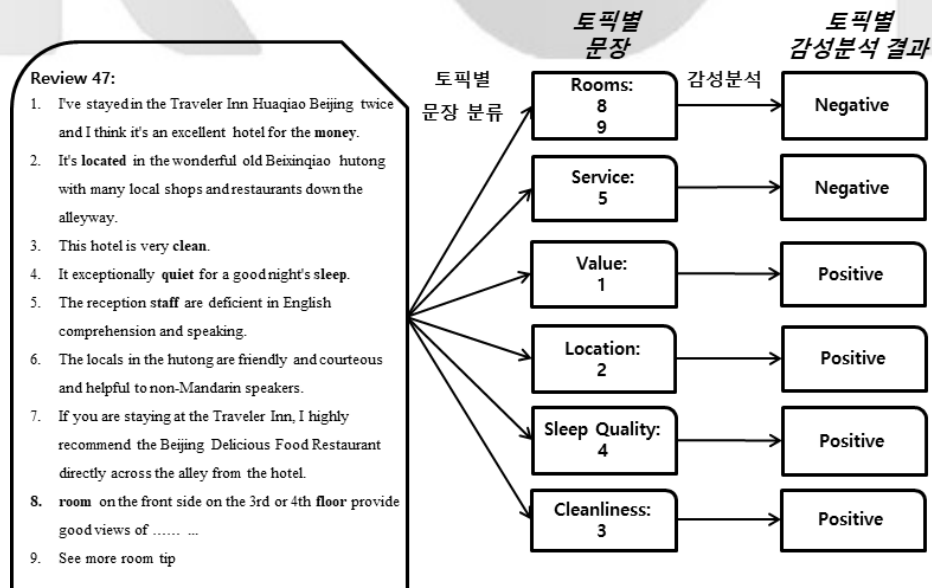
- (1) Review#758에서 포함된 단어들 {단어1, 단어2, 단어3, ... 단어24}에 대해서 감성사전을 이용한 감성분석 실행한 결과에 대해 감성단어별 빈도수 계산

(2) 부정적 단어 \Rightarrow 8개 $\times (-1) = -8$

(3) 긍정적 단어 \Rightarrow 3개 $\times (1) = 3$

4.3.2 토픽별 문장 분류와 감성분석 과정

토픽별 감성분석 과정을 설명하기 위해 리뷰에 대한 토픽별 문장 분류와 감성분석 과정을 <그림 7>과 같이 나타내었다. 자세히 소개하면 이하 두 단계가 있다. 먼저 수집된 각 리뷰를 문장으로 나누고 추출된 토픽별 키워드를 포함한 문장을 토픽별 문장으로 분류한다. <그림 7>에서 볼드체로 표시된 단어들은 추출된 토픽별 키워드이다. 다음으로 토픽별 문장에 대해 어휘 기반 접근법(Lexicon-based Approach)을 이용하여 감성분석을 진행한다. 수집한 데이터 중에서 47번째 리뷰에 대해 토픽별 감성분석 과정을 설계한 것으로, 예를 들어 토픽별 키워드와 대응시켜서 토픽 room과 관련된 문장들이 8번째, 9번째이며, 문장들의 감성분석 결과는 고객이 호텔의 room에 대해 평가할 때 가지고 있는 감성이다. 결과를 보면 room에 대한 감성분석 결과는 부정적이라 할 수 있다. 그 외에 service,



<그림 7> 리뷰에서 문장 분류와 감성분석 과정

value, location, sleep quality, cleanliness에 대한 감성분석 또한 같은 과정으로 진행되었다. 이렇게 각 리뷰에 대해 문장 분류와 감성분석 과정을 반복적으로 진행하여 토픽별 감성분석 결과를 얻을 수 있었다. 여기서 감성사전을 이용해서 감성분석을 할 때 한계점을 생각할 수 있는데, 사전적 의미에 기초한 감성사전의 경우 언어가 사용되는 상황이나 환경에 대한 이해 없이 일반적인 의미에만 의존한다는 것이다(김재봉과 김형중, 2017). 이는 향후 연구에서 상황에 맞는 맞춤형 감성사전 구축을 통해 개선할 필요성이 있다.

4.3.3 토픽별 감성분석 결과

LDA를 통해 추출된 키워드를 바탕으로 6개 상위 토픽과 관련된 최종 키워드를 찾고, 이 키워드들을 포함한 각각 리뷰의 문장을 매칭하여 추출한 토픽별 문장에 대한 감성분석 결과는 <표 6>과 같다. 결과를 보고 각 리뷰의 토픽별 감성분석 결과를 부정 또는 긍정의 감성으로 표현하는데, 이때 평가 리뷰에서 해당 토픽에 대한 문장이 없는 경우는 “NA”로 표기된다. 제안모형을 이용하여 탐색된 각 리뷰의 토픽별 감성분석의 결과를 보면 각 고객의 심리를 파악할 수 있다. 예를 들어 리뷰 1에 대한 결과를

보면 고객이 호텔에 대해 평가하면서 호텔의 Rooms, Location, Cleanliness 등의 토픽에 대해서는 긍정적인 감성을 가지고 있지만 Service, Value 토픽에 대해서는 부정적인 감성을 가지고 있다.

4.4 토픽별 감성분석 모형 검증

본 연구는 각 리뷰의 토픽별로 감성분석을 진행하였으며, 설계된 연구 방법을 테스트하기 위해 속성별 평가점수와 감성분석 결과를 비교해야 한다. 평가점수와 감성값은 부정이나 긍정으로 분류되어 <표 7>과 같은 분류행렬을 통해 비교할 수 있다. 성능 평가 척도로는 정확도(accuracy), 정밀도(precision), 재현율(recall), 그리고 F-1 measure를 사용하였다. 여기서 정확도는 긍정과 부정에 대해 올바르게 분류하는 비율이다. 정밀도는 긍정이라고 예측한 결과에서 실제 긍정에 해당하는 것의 비율이며, 재현율은 실제 긍정 즉, 정답 집합에서 긍정이라고 예측한 것의 비율이다.

$$\text{정확도: } accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$\text{정밀도: } precision = \frac{TP}{TP + FP} \quad (3)$$

<표 6> 리뷰의 토픽별 감성분석 결과

	Rooms	Service	Value	Location	Sleep Quality	Cleanliness
리뷰 1	긍정	부정	부정	긍정	NA	긍정
리뷰 2	부정	부정	부정	NA	긍정	부정
...
리뷰 351,000	긍정	긍정	NA	긍정	NA	NA

<표 7> 분류 행렬

실제 클래스	예측 클래스	
	긍정 C1	부정 C2
긍정 C1	N1,1=올바르게 분류된 C1 데이터 수 (TP: True Positive)	N1,2=C1로 잘못 분류된 C2 데이터 수 (FN: False Negative)
부정 C2	N2,1=C2로 잘못 분류된 C1 데이터 수 (FP: False Positive)	N2,2=올바르게 분류된 C2 데이터 수 (TN: True Negative)

<표 8> 토픽별 감성분석 검증 결과

Dimension	Accuracy	Precision	Recall	F-1 measure
rooms	93.72%	96.87%	96.48%	96.68%
service	96.37%	98.07%	97.12%	97.59%
value	89.35%	93.52%	94.99%	94.25%
location	93.70%	95.17%	98.34%	96.73%
sleep quality	82.91%	84.91%	96.56%	90.36%
cleanliness	95.83%	98.54%	97.54%	97.84%

$$\text{재현율: } recall = \frac{TP}{TP + FN} \quad (4)$$

$$\text{F-1 measure} = \frac{2 \times (precision \times recall)}{precision + recall} \quad (5)$$

제안모형의 성과를 측정하기 위해 고객이 호텔의 속성별로 부여한 평가점수와 본 연구에서 분석한 감성분석 결과를 비교하여 제안모형의 성과를 분석하였다. 분류행렬과 관련된 식을 통해서 정확도(accuracy), 정밀도(precision), 재현율(recall), 그리고 F-1 measure를 계산하여 본 연구에서 구축한 토픽별 감성분석 모형의 성능을 검증한 결과는 <표 8>과 같다.

실험 결과 <표 8>을 보면 sleep quality의 경우 정확도 82.91%, 정밀도 84.91%로, 재현율 96.56%, F-1 척도 90.36%로 다른 토픽에 비해

상대적으로 낮은 값을 나타내고 있다. sleep quality와 value를 제외하면 rooms, service, location, cleanliness의 정확도는 모두 93.70% 이상으로 나타난다. 전체 토픽에 대해 F-1 척도의 최소값은 90.36%이며, 각각의 리뷰에 대한 토픽별 감성분류를 하는 데 매우 효과적이라는 사실을 알 수 있다. 특히 재현율은 6개 상위토픽들이 모두 94.99% 이상의 결과값을 보여주었다.

V. 연구결과 및 향후 연구과제

본 연구는 비구조적인 문서에서 단어들 간 관련성에 따라 분류하는 확률 모델인 LDA를

사용하여 고객들이 직접 방문하고 경험했던 호텔에 남긴 온라인 평가 리뷰 데이터를 대상으로 토픽을 추출하여 분석하였다. 각 토픽에서 빈도가 높은 10개의 키워드를 활용하여 토픽에 대한 해석을 진행하였으며, 추출한 토픽들을 잘 파악하여 호텔에서 고객이 가장 관심이 있는 토픽들이 무엇인지를 탐색하였다. 연구에서 제안한 연구모형의 성능을 검증하기 위해 추출한 토픽들은 리뷰별로 토픽별 감성분석을 실시하고 제안모형의 성능을 검증하기 위해 성능 평가 척도로 재현율, 정밀도, F-1 measure와 함께 결과를 비교하였다. 결과적으로 본 연구에서 제시한 연구 모형이 전체적으로 가장 높은 성능을 보임을 확인하였다.

비즈니스에서 고객들이 경험한 서비스에 대한 평가 정보를 통해 고객들이 중요하게 생각하고 있는 것들이 무엇인지 탐색하는 일이 매우 중요한 일이 되었다. 그러나 소셜네트워크의 사용자가 급증하면서 고객리뷰와 같이 고객들로부터 창출되는 평가 정보의 양 또한 폭발적으로 증가하고 있는 상황에서 비즈니스 관리자가 고객들이 서비스에 대해 가지는 감정이 어떤 것인지 찾아 활용한다는 것이 쉬운 일은 아니다. 그러나 본 연구의 제안모형을 활용하면 관리자는 리뷰별 중요 토픽에 관련된 고객의 감성을 쉽게 탐색할 수 있다. 즉 여행자의 온라인 리뷰를 대상으로 고객의 세부 감성을 다양한 토픽별로 분석할 수 있어 고객에게 여러 속성별로 리뷰작성을 요구하지 않고도 고객의 호텔 속성별 감성을 분석할 수 있으므로, 고객들로 하여금 번거로움을 줄이게 하는 반면, 관리자가 다양한 고객 감성 정보를 활용하여 전략적으로 활용할 수 있도록 할 수 있는 것이다.

그리고 본 연구의 제안모형은 호텔뿐만 아니라 다른 분야에서도 확장하여 적용될 수 있으며, 이를 통해 비즈니스 전략에 있어서 유용하게 활용할 수 있을 것이다.

본 연구의 한계는 다음과 같다. 첫째는 연구에 사용된 고객 리뷰의 언어와 관련된 것으로 본 연구에서는 다른 언어는 고려하지 않고 영어 리뷰만을 사용하여 제안모형을 만들고 분석 및 평가를 수행하였는데, 모형의 일반화를 고려할 때 다른 언어 기반의 리뷰 또한 분석되어야 의미가 있을 것으로 생각된다. 둘째는 감성사전에 대한 것으로 사전적 의미에 기초한 감성사전은 언어가 사용되는 상황이나 환경에 대한 이해 없이 일반적인 의미에만 의존한다는 한계를 가지고 있다. 따라서 연구 대상에 적합한 맞춤형 감성사전을 만들어 활용하는 방법도 모색되어야 할 것이며, 사전적 의미에 기초한 범용 감성사전과 분석결과를 비교하는 작업도 필요할 것이다. 또한 감성분석에서 “중립” 의견을 포함할 수 있는 감성모형 구축과 함께 토픽별 감성분석의 정확도를 높이기 위한 노력도 계속되어야 할 것이다. 본 연구에서는 감성값의 수치를 이용하여 긍정이나 부정에 대한 이진분류만을 고려하였으나, 향후 연구에서는 언어로 표현되는 감성에 대해 극단적인 경우만을 고려하는 것에서 더 나아가 감성의 강도를 고려한 다분류 감성분석 연구를 진행하고자 한다.

참고문헌

김상겸, 장성용, “토픽모델링을 이용한 국내 산업경영공학 연구동향 분석,” 한국경영

- 공학회지, 제21권, 제3호, 2016, pp. 71-95.
- 김재봉, 김형중, “주가지수 방향성 예측을 위한 도메인 맞춤형 감성사전 구축방안,” 한국디지털콘텐츠학회 논문지, 제18권, 제3호, 2017, pp. 585-592.
- 김진화, 변현수, 이승훈, “온라인 리뷰를 활용한 사용자 이해 및 서비스 가치 증대,” 정보시스템연구, 제20권, 제2호, 2011, pp. 21-36.
- 사공원, 하성호, 박경배, “온라인 후기에 내재된 고객의 감성분석과 LQI 차원별 호텔 서비스 품질 평가,” 정보시스템연구, 제25권, 제3호, 2016, pp. 217-245.
- Blei, D. M., “Probabilistic Topic Models,” *Communications of the ACM*, Vol. 55, No. 4, 2012, pp. 77-84.
- Blei, D. M., Ng, A. Y., and Jordan, M. I., “Latent dirichlet allocation,” *Journal of machine Learning research*, Vol. 3, Jan, 2003, pp. 993-1022.
- Bravo-Marquez, F., Mendoza, M., and Poblete, B., “Meta-level sentiment models for big social data analysis,” *Knowledge-Based Systems*, Vol. 69, 2014, pp. 86-99.
- Cao, J., Xia, T., Li, J., Zhang, Y., and Tang, S., “A density-based method for adaptive LDA model selection,” *Neurocomputing*, Vol. 72, No. 7, 2009, pp. 1775-1781.
- Duan, W., Yu, Y., Cao, Q., and Levy, S., “Exploring the impact of social media on hotel service performance: A sentimental analysis approach,” *Cornell Hospitality Quarterly*, Vol. 57, Vol. 3, pp. 282-296.
- Gao, S., Li, X., Yu, Z., Qin, Y., and Zhang, Y., “Combining paper cooperative network and topic model for expert topic analysis and extraction,” *Neurocomputing*, Vol. 257, No. 27, 2017, pp. 136-143.
- Gretzel, U., and Yoo, K. H., “Use and impact of online travel reviews,” *Information and communication technologies in tourism*, 2008, pp. 35-46.
- Hu, M., and Liu, B., “Mining and summarizing customer reviews,” *In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 2004, pp. 168-177.
- Keshavarz, H., and Abadeh, M. S., “ALGA: Adaptive lexicon learning using genetic algorithm for sentiment analysis of microblogs,” *Knowledge-Based Systems*, Vol. 122, 2017, pp. 1-16.
- Kim, H. D., and Zhai, C., “Generating comparative summaries of contradictory opinions in text,” *In Proceedings of the 18th ACM conference on Information and knowledge management*, 2009, pp. 385-394.
- Li, F., Huang, M., and Zhu, X., “Sentiment Analysis with Global Topics and Local Dependency,” *In AAAI*, Vol. 10, July,

- 2010, pp. 1371-1376.
- Lin, C., and He, Y., "Joint sentiment/topic model for sentiment analysis," *In Proceedings of the 18th ACM conference on Information and knowledge management*, 2009, pp. 375-384.
- Litvin, S. W., Goldsmith, R. E., and Pan, B., "Electronic word-of-mouth in hospitality and tourism management," *Tourism management*, Vol. 29, No.3, 2008, 458-468.
- Liu, B., "Sentiment analysis and opinion mining," *Synthesis lectures on human language technologies*, Vol. 5, No. 1, 2012, pp. 1-167.
- Liu, B., *Web data mining: exploring hyperlinks, contents, and usage data*, Springer Science and Business Media, 2007.
- Lu, B., Ott, M., Cardie, C., and Tsou, B. K., "Multi-aspect sentiment analysis with topic models," *In Data Mining Workshops (ICDMW), IEEE 11th International Conference*, 2011, pp. 81-88.
- Marrese-Taylor, E., Velasquez, J. D., and Bravo-Marquez, F., "A novel deterministic approach for aspect-based opinion mining in tourism products reviews," *Expert Systems with Applications*, Vol. 41, No. 17, 2014, pp. 7764-7775.
- Medhat, W., Hassan, A., and Korashy, H., "Sentiment analysis algorithms and applications: A survey," *Ain Shams Engineering Journal*, Vol. 5, No. 4, 2014, pp. 1093-1113.
- Mukherjee, A., and Liu, B., "Aspect extraction through semi-supervised modeling," *In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers*, Vol. 1, 2012, pp. 339-348.
- Nguyen, T. H., Shirai, K., and Velcin, J., "Sentiment analysis on social media for stock movement prediction," *Expert Systems with Applications*, Vol. 42, No. 24, 2015, pp. 9603-9611.
- Pang, B., and Lee, L., "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, Vol. 2, No. 1-2, 2008, pp. 1-135.
- Ravi, K., and Ravi, V., "A survey on opinion mining and sentiment analysis: Tasks, approaches and applications," *Knowledge-Based Systems*, Vol. 89, 2015, pp. 14-46.
- Ren, G., and Hong, T., "Investigating Online Destination Images Using a Topic-Based Sentiment Analysis Approach," *Sustainability*, Vol. 9, No. 10, 2017, pp. 1-19.
- Rhee, H. T., and Yang, S. B., "How does hotel attribute importance vary among different travelers? An exploratory case study based on a conjoint analysis,"

- Electronic markets*, Vol. 25, No. 3, 2015, pp. 211-226.
- Shi, Z., Lee, G. M., and Whinston, A. B., "Toward a better measure of business proximity: Topic modeling for industry intelligence," *MIS Quarterly*, Vol. 40, No. 4, 2015, pp. 1035-1056.
- Song, M., and Kim, S. Y., "Detecting the knowledge structure of bioinformatics by mining full-text collections," *Scientometrics*, Vol. 96, No. 1, 2013, 183-201.
- Stringam, B. B., and Gerdes, J. Jr., "An analysis of word-of-mouth ratings and guest comments of online hotel distribution sites," *Journal of Hospitality Marketing and Management*, Vol. 19, No. 7, 2010, pp. 773 - 796.
- Su, Q., Xu, X., Guo, H., Guo, Z., Wu, X., Zhang, X., and Su, Z., "Hidden sentiment association in chinese web opinion mining," *In Proceedings of the 17th international conference on World Wide Web. ACM*, 2008, pp. 959-968.
- Titov, I., and McDonald, R. T., "A Joint Model of Text and Aspect Ratings for Sentiment Summarization," *ACL*, Vol. 8, June, 2008, pp. 308-316.
- Wang, H., Zhang, D., Zhai, C., "Structural topic model for latent topical structure analysis," *ACL*, 2011, pp.1526-1535.
- Xianghua, F., Guo, L., Yanyan, G., and Zhiqiang, W., "Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon," *Knowledge-Based Systems*, Vol. 37, 2013, pp. 186-195.
- Yu, H., Hatzivassiloglou, V., "Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences," *In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2003.
- Zhao, W., Chen, J. J., Perkins, R., Liu, Z., Ge, W., Ding, Y., and Zou, W., "A heuristic approach to determine an appropriate number of topics in topic modeling," *In proceedings of the 12th Annual MCBIOS Conference*, 2015.
- Zhu, F., and Zhang, X., "Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics," *Journal of marketing*, Vol. 74, No. 2, 2010, pp. 133-148.

홍 태 호 (Hong, Tae-Ho)



KAIST에서 산업공학사를 취득하였고 경영정보시스템을 전공하여 공학석사와 박사 학위를 취득하였다. 현재 부산대학교 경영학과 교수로 재직하고 있다. 딜로이트컨설팅에서 컨설턴트로 재직했으며, 주요 연구관심분야는 비즈니스 애널리틱스, 데이터 마이닝, 오피니언 마이닝, 고객관계관리, 지식 발굴 및 경영 등이다.

니우한잉 (Niu, Hanying)



부산대학교에서 경영학 석사학위를 취득하였다. 주요 관심분야는 데이터 마이닝, 오피니언 마이닝 등이다.

임 강 (Ren, Gang)



부산대학교에서 경영학 석사학위를 취득하였다. 현재 부산대학교 경영학과 박사과정에 재학하고 있으며, 주요 관심분야는 데이터 마이닝, 오피니언 마이닝, 빅데이터 분석, 소셜 미디어, eWOM 등이다.

박 지 영 (Park, Ji-Young)



부산대학교에서 통계학을 전공하여 이학사를 취득하였고, 경영정보, 생산관리 전공으로 경영학 석사 및 박사학위를 취득하였다. 현재 국민대학교 비즈니스 IT 전문대학원 BK21 플러스 사업팀에서 계약교수로 재직중이다. 연구 관심분야는 데이터 마이닝, 오피니언 마이닝, 소셜 네트워크, 고객관계관리 등이다.

<Abstract>

Multi-Topic Sentiment Analysis using LDA for Online Review

Hong, Tae-Ho · Niu, Hanying · Ren, Gang · Park, Ji-Young

Purpose

There is much information in customer reviews, but finding key information in many texts is not easy. Business decision makers need a model to solve this problem. In this study we propose a multi-topic sentiment analysis approach using Latent Dirichlet Allocation (LDA) for user-generated contents (UGC).

Design/methodology/approach

In this paper, we collected a total of 104,039 hotel reviews in seven of the world's top tourist destinations from TripAdvisor (www.tripadvisor.com) and extracted 30 topics related to the hotel from all customer reviews using the LDA model. Six major dimensions (value, cleanliness, rooms, service, location, and sleep quality) were selected from the 30 extracted topics. To analyze data, we employed R language.

Findings

This study contributes to propose a lexicon-based sentiment analysis approach for the keywords-embedded sentences related to the six dimensions within a review. The performance of the proposed model was evaluated by comparing the sentiment analysis results of each topic with the real attribute ratings provided by the platform. The results show its outperformance, with a high ratio of accuracy and recall. Through our proposed model, it is expected to analyze the customers' sentiments over different topics for those reviews with an absence of the detailed attribute ratings.

Keyword: User generated content, Multi-topic, Sentiment analysis, Customer reviews, Latent Dirichlet Allocation

* 이 논문은 2018년 3월 6일 접수, 2018년 3월 23일 게재 확정되었습니다.