붙임2

연구논문/작품 중간보고서

2018 학년도 제 1 학기

제목	감성분석 기반 호텔 리뷰의 극성 분석 및 유저의 선호도 반영 시스템	논문(○) 작품( ) ※해당란 체크
GitHub URL	https://github.com/Ha-yeong/GraduationPaper	
평가등급	지도교수 수정보완 사항	팀원 명단
A	○ 제안기법의 성능 평가가 필요함 ○ 그림을 통한 제안기법 설명이 요구됨 ○	심하영 (학번: 2014312406)

2018 년 3 월 23 일

지도교수 : 김 응 모 서명

■ 요약

인터넷을 통해 정보를 쉽게 공유하게 되면서 소비자들은 제품이나 서비스를 이용하기 전 효율적인 의사 결정을 위해 먼저 작성된 다른 사람들의 의견을 참고한다. 또한 기업들은 이러한 소비자들의 의견을 수집하여 제품의 피드백이나 마케팅에 활용하는 등 비즈니스적인 측면으로 활용한다. 감성분석은 텍스트에 내포된 감성을 식별할 수 있다는 점에서 소비자와 기업 모두에게 주목받고 있는 기술이다. 하지만 아직까지 많은 리뷰 사이트에서 사용자가 제공받을 수 있는 제품/서비스의 순위 정보는 '별점순', '가격순' 등의 일차원적인 데이터를 바탕으로 하고 있다. 본 연구에서는 호텔 예약 사이트인 트립 어드바이저의 서울에 위치한 호텔 100군데에 대해 50개씩의 후기를 수집하였다. 수집한 텍스트 형식의 리뷰를 활용하여 사용자가 호텔을 선택할 때 고려하는 다양한 특성인 '청결도', '서비스', '위치' 등의 정보에 대해 각각의 극성을 분류하고, 특성별 가중치를 적용하지 않은 순서와 적용한 순서를 비교한다. 이를 통해 소비자가 선호하는 호텔의 특성에 가중치를 부여하여 내린 순위가 합리적인 의사 결정에 도움이 된다는 결론을 내릴 수 있다. 나아가 검증된 결론을 근거로 호텔뿐만 아니라 다양한 제품/서비스에 이를 적용시킨 결과를 제공하여 소비자들의 결정을 도울 수 있다는 점에서 연구의 의의가 있다.

■ 서론

감성 분석(Sentiment Analysis)이라고도 불리는 오피니언 마이닝(Opinion Mining)은 제품, 서비스, 조직, 개인, 이슈, 사건, 토픽 등을 다룬 텍스트에 나타난 사람들의 의견, 감성, 평가, 태도, 감성 등을 분석하는 기술을 말한다[1]. 기술의 발전으로 인터넷을 통해 정보를 쉽게 공유하게 되면서 많은 사람들이 제품이나 서비스를 이용하기 전 효율적인 의사 결정을 위해 다른 사람들의 의견을 참고하게 되었다. 또한 기업들은 소셜 네트워크 서비스에 올라온 소비자들의 의견을 수집하여 제품의 피드백이나 마케팅에 활용한다. 이는 기존의 설문조사와 같은 방법에 비해 소비자들의 꾸밈없는 평가를 얻을 수 있고, 시간과 비용을 절약할 수 있다는 점에서 기업에게 유용하다. 따라서 방대한 양의 데이터 속에서 소비자와 기업 모두에게 유용한 정보를 이끌어낼 수 있는 오피니언 마이닝이 주목받게 되었다.

오피니언 마이닝의 상위 분류인 텍스트 마이닝(Text Mining)과의 차이는 다음과 같다. 텍스트 마이닝은 텍스트에 내포된 '사실'에 초점을 두고 있으나, 오피니언 마이닝은 텍스트에서 사용자가 취하는 '태도'에 집중한다는 점이다[2]. 사용자가 문서 전체, 문장, 혹은 텍스트가 다루고 있는 특성 개체의 속성에 대해 긍정과 중립, 부정 중 어떤 입장을 취하고 있는지는, 각 단어별로 감성 값이 부여되어 있는 감성 사전을 사용해서 극성을 계산하는 과정을 통해 알아낼 수 있다.

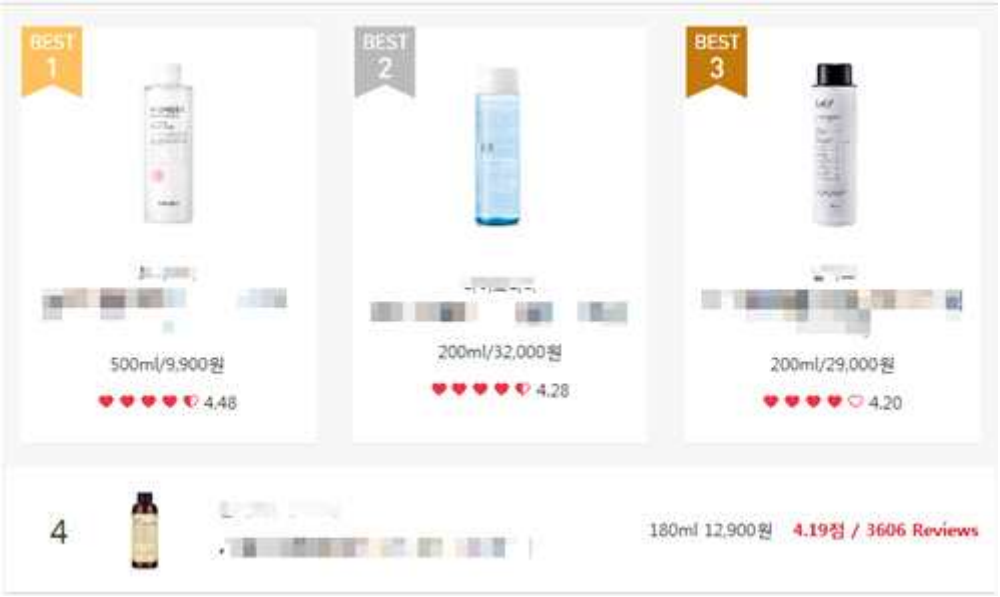
오피니언 마이닝의 단계는 크게 1) 데이터 수집 2) 전처리 3) 감성 사전의 구축 4) 극성 분석의 네 가지 단계로 이루어진다. 이 때 단어의 극성, 즉 긍정/부정을 판별하기 위해서는 SentiWordNet과 같이 이미 만들어져 있는 감성 사전을 참조하거나(Dictionary-based), 데이터에서 추출한 말뭉치를 기반으로 감성 사전을 구축하여 사용한다(Corpus-based).

첫 번째 방법인 사전 기반 감성 사전은 각기 다른 제품 혹은 서비스의 특징을 반영하지 못한다는 단점이 있다. 예를 들어 '빠르다'라는 서술어는 '서비스 응대가 빠르다'의 경우 긍정적으로 쓰이지만 '배터리 소모가 빠르다'에서는 부정적인 서술어로 쓰여야 한다. 이러한 부분을 보완하기 위해 두 번째 방법인 말뭉치 기반 감성 사전을 사용할 수 있다. 말뭉치 기반 감성 사전은 높은 별점의 후기에서 빈번하게 나오는 단어에는 긍정적인 값을, 낮은 별점의 후기에서 빈번하게 나오는 단어에는 부정적인 값을 매기는 방법을 통해 구축한다. [3]에서는 제품 특징에 따라 서술어의 의미 방향이 다르게 사용되는 것을 고려하여 말뭉치 기반 감성 사전을 구축하였고, 그 결과 서술어를 일반적인 의미 방향으로 분류한 감성사전을 사용했을 때보다 더 좋은 성능을 보임을 밝혔다.

이러한 점에 착안하여 제품이나 서비스에 따라 특징이 되는 키워드를 추출하고, 말뭉치 기반 감성 사전을 활용하여 키워드 각각에 대한 리뷰의 극성을 판별해볼 수 있다면 의미 있는 정보를 얻을 수 있겠다고 생각했다. 우선 전자제품, 화장품, 호텔 후기를 볼 수 있는 대표적인 사이트들을 방문하여 사용자에게 보여지는 정보를 확인해보았다.

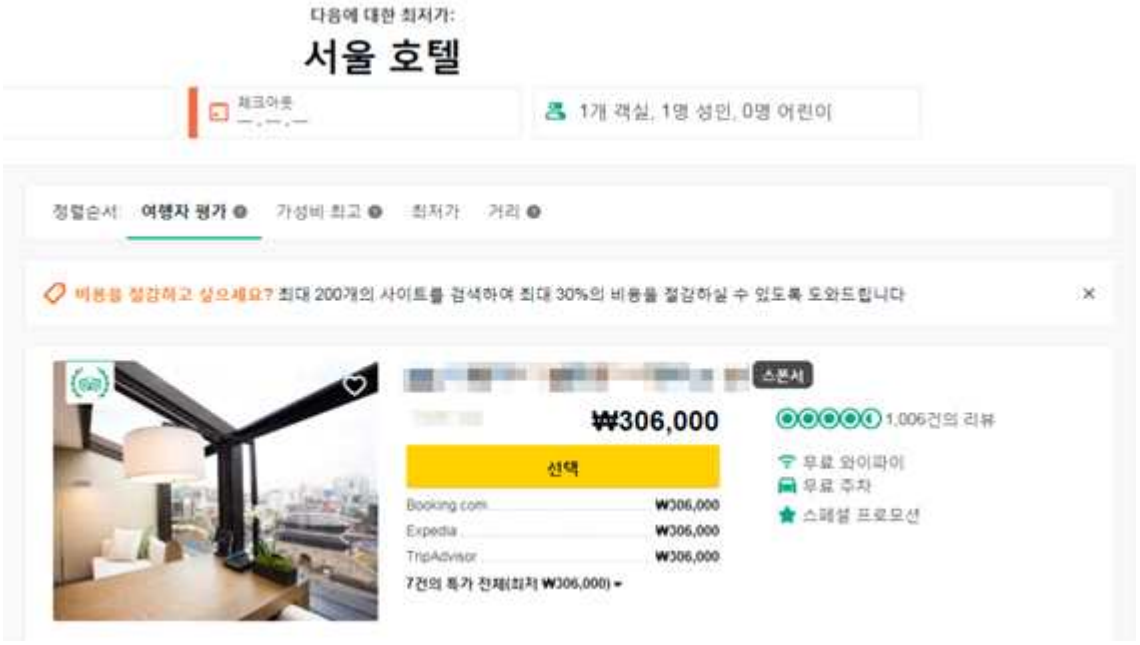
(그림 1) 11번가 노트북을 랭킹순으로 조회한 결과

(그림 1)은 11번가에서 특정 브랜드의 노트북을 랭킹순으로 조회한 결과이다. 기본적으로 조회되는 '11번가 랭킹순'은 광고가 포함되어 있어 소비자들이 합리적인 결정을 하는 것에 영향을 끼칠 수 있다. 그 외에 '11번가 인기순', '낮은 가격순', '누적 판매순', '많은 리뷰순', '높은 평가순', '높은 가격순' 등으로 조회를 할 수 있어 소비자가 얻을 수 있는 정보는 가격, 별점, 판매 순위이다.



(그림 2) 글로우픽에서 스킨케어제품을 조회한 결과

(그림 2)는 화장품 리뷰 전문 사이트인 글로우픽에서 스킨케어제품을 조회한 결과이다. 사용자는 한 눈에 별점을 확인할 수 있지만, 수분감이나 유지 시간, 트러블 유발 여부 등 사용자에 따라 고려하는 부분이 다양한 화장품의 특성상 높은 별점만 보고 선택을 하기는 힘들어 다른 사용자들의 후기를 일일이 직접 읽어야 한다. 화장품의 종류에 따라 중요하게 여겨지는 항목들의 별점을 각각 순서대로 조회할 수 있다면 사용자는 수많은 제품들 사이에서 좀 더 빠르고 편한 의사결정을 할 수 있을 것이다.



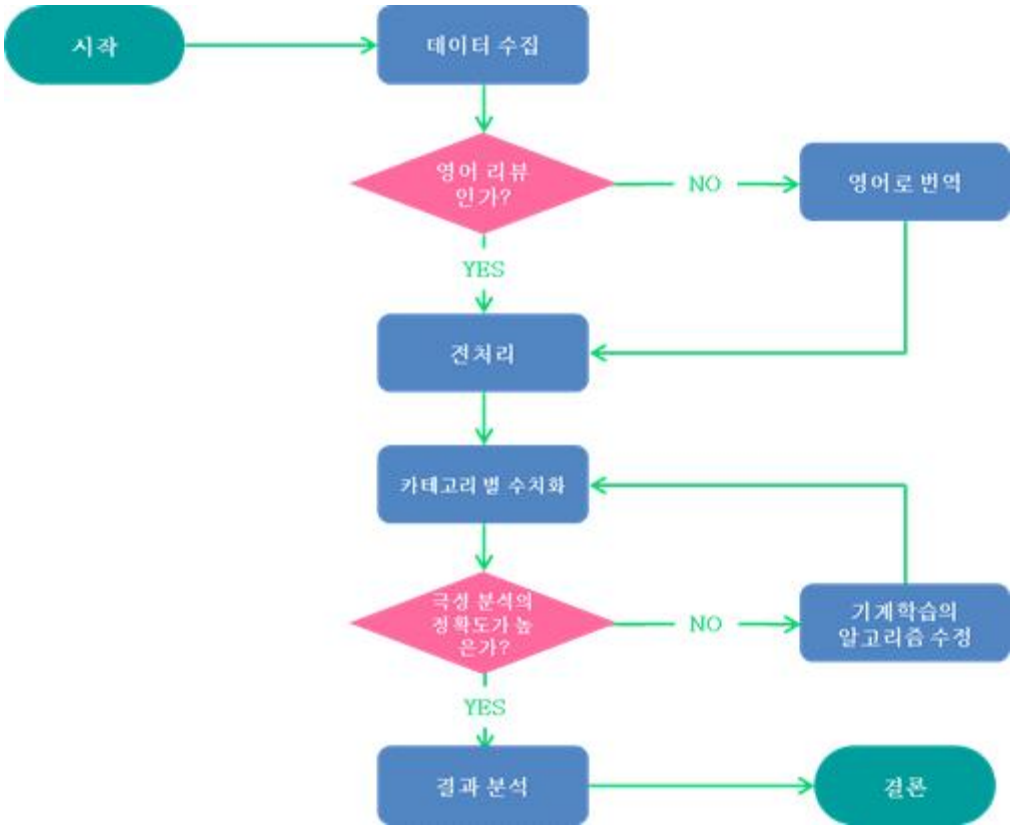
(그림 3) 트립 어드바이저 서울의 호텔을 조회한 결과 1

(그림 4) 트립 어드바이저 서울의 호텔을 조회한 결과 2

(그림 3)은 호텔, 음식점 등을 예약하거나 후기를 공유하는 트립 어드바이저에서 서울의 호텔을 조회한 결과이고, (그림 4)와 같이 각 호텔의 별점이 보여진다. 사용자는 여행자 평가(별점), 가성비, 가격, 중심부로부터의 거리 순으로 호텔을 조회할 수 있다. 호텔을 예약할 때 우선 위치의 편리성이나 가격이 크게 고려되기 때문에 가격과 거리 순서에 따른 호텔 정보가 제공되는 것은 편리한 부분이다. 하지만 그런 객관적인 사실 이외에 위생 상태나 서비스, 편의 시설 등에 대한 주관적인 후기들도 사용자가 편리하게 확인할 수 있다면 호텔을 선택하기 수월할 것이고 실제 사용한 후에도 만족도가 높을 것이다.

앞서 제시한 세 가지 사례를 통해 쇼핑몰이나 예약 사이트에서 제품 또는 서비스를 제공하는 형태를 확인해보았다. 가격, (호텔의 경우) 거리와 같은 객관적인 정보 외에 쉽게 확인할 수 있는 정보는 다른 사용자들이 남긴 별점이다. 하지만 제품 또는 서비스를 선택할 때 가장 중시하는 부분이 사용자마다 다르기 때문에 합리적인 선택을 하기 위해서는 다른 사용자들이 남긴 후기를 일일이 읽어보아야 한다. 이는 선택을 하는데 걸리는 시간과 노력을 증가시키고, 몇몇의 극단적인 후기들은 합리적인 선택에 있어서 지장을 주기도 한다.

이러한 문제를 해결하기 위해서는 애초에 별점을 입력받을 때 제품 또는 서비스별로 중요한 항목별로 별점을 매기는 방법이 있다. 다른 방법으로는 사용자들이 남긴 수많은 텍스트 형식의 후기들을 수집하고 항목별로 오피니언 마이닝을 수행하는 방법이 있다. 이 방법은 주관적인 후기를 수치화하여 사용자에게 제시할 수 있으므로 다른 사용자들이 남긴 후기를 일일이 읽어보는 수고로움을 덜어줄 수 있을 것이다. 본 연구에서는 두 번째 방법을 사용해보았다.



(그림 5) 논문 구성

따라서 이 연구의 목적은 특정 제품 또는 서비스의 텍스트 형식의 후기들을 키워드별로 나누어서 오피니언 마이닝으로 수치화시켰을 때 사용자가 합리적인 선택을 하는 것에 도움을 주는지 확인해보는 것이다. (그림 5)와 같이 이 연구는 데이터 수집, 수집한 데이터에 대한 전처리, 기계 학습을 이용한 카테고리별 수치화, 결과 분석의 단계로 수행될 것이다. 카테고리를 수치화한 후에는 각 항목마다의 가중치를 다르게 두고, 그 결과가 일반적인 '별점 순위'와 달라지는지를 확인할 것이다. 결과가 예상대로 도출된다면 사용자가 합리적인 선택을 하는 것에 대해 키워드별로 수치화된 후기가 도움이 된다는 것을 증명할 수 있다.

이상과 같이 1장에서는 연구의 배경과 목적에 대한 소개를 하였다. 2장에서는 연구 주제와 관련된 논문에 대해 설명할 것이다. 3장에서는 각 단계의 연구를 구현한 구체적인 방법에 대해 소개할 것이다. 4장에서는 결과를 예측하고 오피니언 마이닝의 정확도를 계산하는 방법을 다룰 것이다. 5장에서는 연구를 진행하며 느낀 소감을 간단히 다루겠다.

■ 관련연구

1. 감성분석 연구의 발전 배경

감성분석 연구는 컴퓨터 과학 분야의 자연어 처리 중 하나의 주제로써 연구되기 시작하여 현재는 다양한 학계 및 산업으로 확장되어 연구되고 있다. 이러한 확장은 제품/서비스에 대한 사람들의 의견을 다양한 소셜 미디어를 통해 얻을 수 있게 되었고, 비즈니스적인 측면에서 대중의 의견을 분석함으로써 이윤을 극대화시키고자 하는 동기가 있었기 때문이다 [4].

감성분석 연구는 2000년대 이후부터 활발히 이루어졌다. [5]에서는 감성분석에 접근하는 다양한 연구 방법들과 이론을 fact-based 분석에서 적용되고 있던 다양한 방법론과 비교하며 정리했다. 감성분석은 본질적으로는 텍스트의 주제를 정치, 과학, 스포츠 등으로 분류하는 것에서 파생되었기 때문이다. 다만 텍스트의 주제 분류는 주제를 나타내는 키워드에 집중하지만, 감성분석은 텍스트에 나타난 감성을 나타내는 단어에 주목한다는 점이 다르다. [6]은 발전된 감성분석 방법론을 적용하는 방법에 대해 다뤘다.

2. 기계학습에 기반을 둔 감성분석

[6]에 따르면, 감성분석은 크게 지도학습(supervised learning)과 비지도학습(unsupervised learning) 기반의 기계학습에 바탕을 두고 있다.

2-1. 지도학습 기반의 기계학습

[4]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. 2002년 Pang과 Lee에 의해 발표된 [7]은 이러한 지도학습 모델을 사용하여 영화 리뷰를 긍정과 부정의 두 가지로 분류한 첫 번째 연구이다. 이 연구에서는 영화 리뷰 평점이 4~5점일 때를 긍정으로, 1~3점일 때를 부정으로 정의했다. 나이브 베이즈, 최대 엔트로피, SVM의 세 가지 학습 분류기를 사용했으며 결과적으로 SVM이 가장 성능이 높음을 밝혔다. 이후 다양한 연구에서 감성의 분류를 긍정/부정뿐만 아니라 긍정/부정/중립으로 나누는 것을 시도해보기도 하고 확률 언어 모델인 n-gram에 변화를 주는 등 다양한 시도를 통해 정확도를 높여나갔다.

[12]에서는 지도학습 방법을 사용한 가우시안 나이브 베이지안 분류기에 반복 알고리즘을 적용하여 감성분석을 더욱 효과적으로 할 수 있도록 하였다. 데이터의 분포를 가우시안으로 가정하고 베이즈룰을 이용하는 가우시안 나이브 베이지안 분류기는 주로 사람의 감정 상태를 분류, 판단하는 연구에서 사용된다. 기존의 방법의 경우 훈련 데이터(Training data)를 이용하여 고정된 분류기를 설계하고, 이 분류기를 검증 집합(Validation data)을 이용하여 평가하는데, 고정된 분류기는 감정과 같이 개개인의 차이로 인해 생기는 다양한 데이터를 분류할 때 문제점이 발생하며 딥러닝이나 AI 분야에 적용하기 어렵다. 따라서 이 연구에서는 기존의 가우시안 나이브 베이지안 분류기에 반복 알고리즘을 적용하여 이를 보안하였다. 제안하는 방법에서는 검증 집합을 이용하여 평가할 때, 실제론 그 부류에 속하는 데이터이지만 오분류 된 데이터를 기억한다. 오분류 데이터 개수가 일정 수준이 넘어가면 이를 기존의 훈련 집합에 포함하여 분류기를 재설계한다. 검증 결과 기존보다 정확도 높은 분류기를 얻었다.

2-2. 비지도학습 기반의 기계학습

감성분석에서 지도학습 기반의 기계학습이 대다수를 차지하지만, [8]과 [9]의 연구과 같이 비지도학습 기반으로 감성분석을 수행한 경우도 있다. [8]에서는 의견을 나타내는데에 빈번하게 쓰이는 구문의 패턴을 품사 태그의 집합으로 분석하였다. 그리고 연속된 단어의 품사가 의견을 나타내는 데에 빈번하게 쓰이는 구문의 패턴에 포함된다면 이들을 추출하는 방법을 사용했다. 구문의 패턴 중 첫 번째는, 형용사인 단어 뒤에 명사가 나오는 경우로 “This piano produces beautiful sounds.”에서 “beautiful sounds”가 이러한 경우의 예시가 될 수 있다. 그 외에 [9]에서는 감성사전을 기반으로 한 방법을 사용했다.

최근에는 이러한 방법 외에도 딥 러닝(deep learning) 혹은 딥 뉴럴 네트워크(deep neural network)에 기반을 두는 감성분석 연구도 증가하고 있다. 본 연구에서는 지도학습을 기반으로 하는 감성 분류기를 사용했고, 주어진 텍스트에 나타난 감성을 긍정/부정으로 분류하는 것을 넘어서서 텍스트에 나타난 제품/서비스의 특성별로 긍정/부정을 분류하는 것을 목표로 했다.

3. 특성별 감성분석

텍스트에 나타난 여러 가지 특성별로 감성분석을 수행하는 것을 Aspect Based Sentiment Analysis(ABSA)라 한다. [10]은 특정 분야의 특징 단어들을 추출하고, 각 특징 단어에 대해 감성 분석을 수행하는 ABSA 시스템을 다음과 같은 세 가지 단계로 나눴다.

첫 번째는 특징 단어 추출(aspect term extraction) 단계로, 특정 도메인에 대한 텍스트를 입력받아 그 문장에서 다뤄지는 단어들을 추출한다. 예를 들어 '컴퓨터' 제품에 대한 리뷰에서 다뤄지는 단어들은 '배터리', '하드 디스크'등이 있을 수 있다. 두 번째 단계는 특징 단어 군집화(aspect term aggregation) 단계로, 첫 번째 단계에서 얻은 특징 단어들을 비슷한 것들끼리 군집화하는 단계이다. 예를 들어 'price'와 'cost'는 하나의 군집으로, 'design'과 'color'는 다른 하나의 군집을 이루게 된다. 세 번째 단계는 특징 단어 극성 예측(aspect term polarity estimation) 단계로, 특징 단어들의 군집에 대해 전체 텍스트의 긍정/부정의 정도를 예측한다.

[11]에서 Maria Pontiki 외 5명의 연구자들은 감성 분석에 관한 많은 연구가 해당 연구에서 다루는 도메인의 특징에 무관하게 일반적인 긍정/부정을 찾아내는 것에서 문제점을 찾았다. 그리고 연구를 통해 음식점과 노트북이라는 도메인의 특징 단어들을 파악하고 각각의 특징 단어에 대해 감성 분석을 수행하였다. 연구는 다음과 같은 네 가지 단계로 수행되었다.

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

예를 들어 "The staffs were not that kind, but the menu was great."라는 문장은 각 단계를 거쳐 다음과 같은 결과를 낸다.

1. aspect term extraction 단계 : {staff, menu}
2. aspect term polarity 단계 : {staff: negative, menu: positive}
3. aspect category detection 단계 : {service, food}
4. aspect category polarity 단계 : {service: negative, food: positive}

첫 번째와 두 번째 단계는 특성 카테고리가 정의되어있지 않은 경우 유용하다. 이런 경우 자주 언급되는 특징 단어들을 카테고리화하여 감성 분석을 수행하면 된다. 세 번째와 네 번째 단계는 특성 카테고리가 정의되어있는 경우 사용한다. 해당 연구에서는 위의 단계를 거쳐 음식점과 노트북에 대한 ABSA benchmark datasets를 XML 형식으로 만들고 배포했다.

■ 제안 작품 소개

이번 연구는 크게 데이터 수집, 수집한 데이터에 대한 전처리, 기계 학습을 이용한 카테고리의 수치화, 결과 분석의 단계로 나누어진다. 이번 장에서는 각 단계의 의의와 구현 단계에 대해 소개하겠다.

1. 데이터 수집 단계

데이터를 수집하는 단계이다. 데이터를 얻기 위한 대상 웹페이지를 결정하는 단계에서 고려한 사항은 다음과 같다.

1) 특정 제품이나 서비스에 대한 후기를 볼 수 있어야 한다.
2) 포스팅 형식보다는 댓글 형식으로 된 후기가 제품/서비스와 관련 없는 내용이 적게 포함되어 있고 데이터 수집 과정이 수월할 것이다.
3) 다양한 의견 추출을 위해 제품/서비스 별 후기가 최소 50개 정도는 있어야 한다.

1장에서와 같이 11번가 노트북, 글로우픽 화장품, 트립 어드바이저 서울 호텔에 대해 조회해본 결과 트립 어드바이저 후기의 길이가 적당하고 카테고리가 명확하였다. 또한 인터넷 용어의 사용 빈도가 낮은 점에서 전처리 단계가 수월한 부분이 기대되었다. 11번가 노트북과 글로우픽 화장품에서는 다양한 브랜드의 제품에 대한 소비자들의 후기를 얻을 수 있었으나 제품 자체와는 상관없는 내용도 후기의 많은 부분을 차지했고 인터넷 용어의 사용 빈도가 높았다.

따라서 트립 어드바이저가 오피니언 마이닝을 수행하기에 가장 적절한 데이터를 제공했다. SPA로 구현된 웹페이지의 특성상 Chrome Web Driver와 Selenium을 사용하여 크롤러를 구현했다. Web Driver는 브라우저에서 제공되는 API로, 이를 이용하면 코드를 통해 실제 사용자가 브라우저를 다루는 것처럼 사용할 수 있다. 트립 어드바이저에서 서울에 위치한 호텔 100곳의 후기를 최근에 작성된 순서대로 50개씩 수집하여 텍스트 파일로 저장하기로 하였다. 이를 위해 크롤러는 크롬 브라우저를 열어 트립 어드바이저에서 각 호텔의 정보가 나와 있는 주소로 이동하고, 리뷰 '더 보기' 버튼을 클릭하여 텍스트 형식의 리뷰를 얻어온 뒤 다음 페이지로 이동하는 과정을 반복한다.

데이터 수집 과정에서 두 가지 문제가 발생하였고 이를 해결했다. 첫 번째 문제는 리뷰 중 상당 부분은 외국어로 작성된 것이었고, 이들을 제외하고 한국어 리뷰만 수집한다면 데이터양이 현저하게 줄어든다는 것이었다. 트립 어드바이저 홈페이지에서 제공되는 자동 번역 기능을 끄고 한국어와 영어로 작성된 리뷰만 수집하였다. 두 번째 문제는 후기가 많이 있는 호텔은 별점이 대체로 4.0과 4.5에 편향되어 있다는 것이었다. 별점은 방문자들이 '아주 좋음', '좋음', '보통', '별로', '최악'으로 평가한 것을 각각 5, 4, 3, 2, 1점씩 부여하여 소수 첫째 자리까지 얻은 결과라는 것을 알아냈다. 따라서 크롤링 과정에서 각 호텔의 별점을 소수 둘째 자리까지 다시 계산하여 별점을 더 세분화하였다.

2. 전처리 단계

텍스트를 전처리하는 방법은 데이터 상황에 따라 가변적이다. 따라서 데이터의 분석 방향에 적합한 전처리 계획을 세우는 것이 중요하다. 전처리를 수행한 데이터는 트립 어드바이저에서 얻은 100개의 호텔에 대한 리뷰이다. 한국어로 작성된 리뷰의 경우 불필요한 한글 자음, 모음, 불필요한 문장 부호 등을 제거해야하고 영어로 작성된 리뷰의 경우 문장 부호로 이루어진 이모티콘과 불필요한 문장 부호를 제거해야한다. Python의 're' 라이브러리를 이용하여 다음과 같은 단계로 전처리 작업을 수행했다. 한국어 리뷰는 전처리 과정을 거친 뒤 구글 번역기를 이용하여 영어로 번역하였다.

1) re.sub('[ㄱ-ㅎㅏ-ㅣ]', '.', line)
정규표현식을 활용하여 한글 자/모음을 마침표로 치환하는 작업이다.
대부분의 문장에서 ㅋㅋ나 ㅎㅎ, ㅠㅠ 등은 문장 끝에 사용하기 때문이다.

2) re.sub('[-_@#$*$(){}:"]', '', line)
내용에 영향을 끼치지 않는 불필요한 문장 부호들을 제거했다.

3) re.sub('[.!~;^]+', '.', line)
마침표, 느낌표, 세미콜론과 같은 기호들이 문장의 끝부분에서 반복적으로 사용되는 경우가 많음을 확인하고 이를 하나의 마침표로 치환했다.

4) re.sub('[&]+', ' and ', line)
기호 &을 and로 치환했다.

5) re.sub('[?]+', '?', line), re.sub('[,]+', ',', line), re.sub('[ ]+', ' ', line)
반복적으로 사용된 물음표, 쉽표, 공백을 하나로 줄였다.

한국어 리뷰를 영어로 번역하는 과정에서 다음과 같은 문제가 발생하였다.

<표 1> 불필요한 줄 바꿈으로 생긴 번역 결과

	Case 1.	Case 2.
리뷰	아침 조식도 꼭 나와야 할 메뉴는 다 있어서 만족스러웠어요.	아침 조식도 꼭 나와야 할 메뉴는 다 있어서 만족스러웠어요.
번역 결과	Breakfast breakfast It was satisfactory	Breakfast was also satisfying because we had all the menus that we had to come out for

<표 1>의 Case 1과 같이 불필요한 줄 바꿈이 있는 경우 구글 번역기를 통해 번역한 결과가 어색해지는 것을 확인하였다. 이 문제를 Case 2와 같이 줄 바꿈을 한 칸의 공백으로 변경하는 것으로 간단히 해결하였다. 문장 끝에 마침표가 붙어있지 않은 경우에도 구글 번역기에서 문장의 끝을 인식해서 번역이 잘 되었고 마침표도 자동으로 붙여주는 것을 확인하였다.

3. 기계 학습을 이용한 카테고리의 수치화 단계

전처리 과정을 마친 리뷰에서 '청결도', '음식', '위치', '가성비' 등의 카테고리 별 긍정/부정을 수치화하는 단계이다. MonkeyLearn에서 제공하는 기계학습 기반의 호텔 기반 긍정/부정 분류 API와 카테고리 분류 API를 사용하여, 카테고리별로 극성을 분석하는 방법을 사용했다. 먼저 두 가지 분류 API는 다음과 같은 방법으로 구현되었다.

3-1. Sentiment Analysis Model with Scrapy and MonkeyLearn

a. Scrapy를 사용하여 트립 어드바이저의 뉴욕 호텔 리뷰를 수집한다.
b. 1에서 얻은 후기들을 training samples로 사용하여 기계 학습 모델링을 수행한다. 별점이 3 초과인 후기는 긍정으로, 3 이하인 후기는 부정으로 정의한다. 긍정과 부정의 후기가 각각 5000개 정도가 될 때까지 리뷰를 수집한다.
c. Monkeylearn에서 custom text classifier를 만들고, 수집한 정보를 사용하여 학습시킨다.
d. 학습이 끝나면 n-gram range, 알고리즘 등의 셋팅을 조절하며 분석의 정확도를 높인다.

위의 과정을 통해 만들어진 긍/부정 분류기는 호텔 리뷰에 특화된 분류기이다. 또한 긍/부정을 분류하는 것에 그치지 않고 probability를 0과 1사이의 값으로 제공하기 때문에 0과 가까운 결과는 중립에 가까운 것으로 판단하여 감성 분석 대상에 포함시키지 않을 수 있을 것이다. MonkeyLearn API를 통해 Python에서 분류기를 사용해보았고 다양한 케이스를 분류해봤을 때 품사의 패턴의 특성을 바탕으로 NLTK와 SentiWordNet을 사용하여 극성을 분류해보았을 때보다 간단하고 정확하였다. 최종적으로 완성된 분류기는 약 90%의 정확도를 가진다.

3-2. Aspect Analysis from reviews using Machine Learning

a. Scrapy를 사용하여 Booking.com의 뉴욕 호텔 리뷰를 수집한다.
b. 텍스트를 문장 단위로 분해하고 문장에서 but이 나타나면 다시 분해한다. "The location was good, but the service was lacking." 위와 같이 but을 전후로 위치에 대해서는 좋은 평가를, 서비스에 대해서는 나쁜 평가를 내리고 있기 때문이다.
c. Amazon Mechanical Turk를 사용하여 2000개의 문장을 'cleanliness', 'comfort' 등으로 분류했다.
d. 이를 SVM(Support Vector Machine)을 사용하여 기계학습 시킨다. 약 80%의 정확도를 가진다.

Amazon Mechanical Turk란, 온라인에 등록된 인력을 할당, 작업에 따라 보수를 지급하는 구조로 이뤄진 일종의 온라인 벼룩시장으로, 1770년 개발된 자동 체스 기계의 이름을 따왔다. 자동 체스 기계는 겉으로는 자동으로 체스를 기계가 두는 것처럼 보이지만 실제로는 내부에 사람이 들어가서 기계를 다루고 있었다고 한다. 기술의 발전에도 불구하고 여전히 인간의 판단력이 요구되는 현장이 많은데, Amazon Mechanical Turk이 서비스 이용자와 작업자를 중개해주는 것이다.

이러한 과정을 통해 만들어진 카테고리 분류기는 호텔 리뷰에 특화된 분류기이다. 또한 긍/부정을 분류하는 것에 그치지 않고 probability를 0과 1사이의 값으로 제공한다. MonkeyLearn API를 통해 Python에서 분류기를 사용할 수 있다.

■구현 및 결과분석

앞서 기계학습을 활용하여 '호텔'에 특화된 긍/부정 분류기와 카테고리 분류기를 구현하는 방법을 다뤘다. 긍/부정 분류기는 주어진 문장을 'Good' 또는 'Bad'로 분류하며 각각의 경우에 대해 0에서 1 사이의 probability 값을 제공한다. 즉 0에 가까운 값일수록 문장이 중립에 가까운 것이다. 알고리즘 구현 과정에서는 'Bad'에 -1을 곱하여 극성 분석 결과가 -1에서 1의 값을 가지도록 하고, -0.7에서 0.7 사이의 값은 중립으로 판단, 그 외의 값만 유의미한 값으로 분류한다.

카테고리 분류기는 주어진 문장을 입력받아 호텔에 관련된 특성 중 어떤 특성을 다루는지를 도출한다. 호텔에 관련된 특성은 'Cleanliness', 'Comfort & Facilities', 'Food', 'Internet', 'Location', 'Staff', 'Value for money'로 나누어진다. 입력받은 문장이 어떤 카테고리에도 속하지 않지만 유의미한 긍정/부정의 값을 가지는 경우 'General'이라는 특성의 극성을 변화시키기로 한다. 카테고리 분류기 또한 긍/부정 분류기와 마찬가지로 각각의 경우에 대해 0에서 1 사이의 probability 값을 제공한다. 0과 가까울수록 정확도가 낮으므로 0.7에서 1 사이의 값만 유의미한 값으로 분류한다.

이러한 분류기들의 특징을 활용하여 다음과 같은 방법으로 알고리즘을 구현하면 호텔에 특화된 Aspect Based Sentiment Analysis(ABSA) 시스템을 구현할 수 있다.

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 line splitted by 'but', 'and', ',')
9.  return sum of array[]
```

array[0]은 'Cleanliness'의 극성을, array[1]는 'Comfort & Facilities'의 극성을 나타낸다. 3번 줄의 경우와 같이 아무런 카테고리에 포함되지 않는 경우 array[7] 즉 'General'의 극성을 더한다. 5번 줄과 같이 문장이 카테고리 하나에 포함되는 경우는 해당 카테고리의 극성을 array에 더한다. 7번 줄과 같이 문장이 여러 개의 카테고리를 다루는 경우는 'but', 'and', ',' 와 같은 단위로 다시 나누어 알고리즘을 수행한다.

예를 들어 다음과 같은 문장이 있다.
"Room was clean, but the staffs were not kind." — (a)

문장 (a)에 대한 2)의 결과, 즉 분류기를 사용하여 카테고리를 판단한 결과는 'Cleanliness'와 'Staff'이다. 각각의 probability는 0.938, 0.913이므로 둘 다 유의미한 카테고리이다. 하지만 그 개수가 두 개 이상이므로 'but'을 기준으로 다시 두 개의 문장으로 분류한다. 첫 번째 문장인 "Room was clean"은 'Cleanliness' 카테고리에 속하고 긍/부정 분류기를 통한 결과는 'Good'으로, 0.86의 probability를 가진다. 두 번째 문장인 'the staffs were not kind."는 'Staff' 카테고리에 속하고 긍/부정 분류기를 통한 결과는 'Bad'로, 0.72의 probability를 가진다. 따라서 전체 문장의 극성은 0.86 - 0.72의 결과인 0.14이다.

위의 방법을 통해 'Cleanliness', 'Comfort & Facilities', 'Food', 'Internet', 'Location', 'Staff', 'Value for money'의 각각의 극성을 구할 수 있고 이를 모두 더해서 후기가 전체적으로 긍정의 양상을 띠는지 부정의 양상의 띠는지를 알아낼 수 있다. 그 결과로 100군데의 호텔에 대해 새롭게 계산된 별점 순서를 얻을 수 있고, 이를 기존에 트립 어드바이저에서 제공한 별점 순서와 비교하여 새롭게 구현된 알고리즘의 정확성을 확인할 수 있다.

새롭게 구한 순위가 트립 어드바이저에서 제공한 별점 순서와 그 양상이 비슷하고 이를 통해 정당성을 확보할 수 있다면, 'Cleanliness', 'Comfort & Facilities', 'Food', 'Internet', 'Location', 'Staff', 'Value for money' 각각의 특성에 대해 다른 가중치를 부여한 결과를 구해본다. 예를 들어 앞에서 언급된 것처럼 문장 (a)는 'Cleanliness'에 대해 0.86의 극성을, 'Staff'에 대해서는 -0.72의 극성을 가진다. 앞에서는 단순히 이들을 더하여 0.14라는 극성을 얻었다. 하지만 호텔을 선택할 때 'Cleanliness'를 'Staff'보다 중요하게 여기는 소비자에게는 'Cleanliness'에 더 높은 가중치를 부여한 결과가 더 도움이 될 것이고, 그 반대의 경우 'Staff'에 더 높은 가중치를 부여한 결과가 더 도움이 될 것이다. 실제로 전자의 경우 0.86 x 2 - 0.72 즉 1의 긍정값을 가지고, 후자의 경우 0.86 - 0.72 x 2 즉 -0.58의 부정값을 가지기 때문에 큰 차이를 보이는 것을 알 수 있다.

정리하면, 앞으로는 Python을 사용하여 위의 알고리즘을 실제로 구현하고, 수집한 데이터에 대해 적용시켜 각각의 특성에 대한 극성값을 구할 것이다. 이를 트립 어드바이저에서 제공한 별점 순서와 비교하며 그 정확도를 높일 것이다. 그리고 각각의 특성에 대해 다른 가중치를 부여하며 이것이 트립 어드바이저의 별점 순서와 다른지를 확인할 것이다.

이를 통해 소비자가 원하는 대로 호텔의 특성에 다른 가중치를 부여하는 것이 의사결정 과정에 있어서 더 효율적이라는 결과를 내릴 수 있을 것이다. 차후에는 이러한 서비스를 호텔뿐만 아니라 다양한 제품과 음식점, 항공사 등으로 확장시켜 관련 사이트에서, 소비자가 우선시하는 특성의 순서대로 제품/서비스를 조회하도록 할 수 있다.

■ 결론 및 소감

본 논문에서는 소비자가 선호하는 호텔의 특성에 가중치를 부여하여 구한 순위가 합리적인 의사 결정 시 도움이 된다는 것을 제안하였다. 웹 크롤러를 사용하여 트립 어드바이저에서 서울에 위치한 호텔 100군데의 텍스트 형식의 리뷰를 수집하고, 전처리 과정을 거쳐 지도학습을 기반으로 학습시킨 기계학습 분류기를 사용하여 호텔의 전체적인 순서를 다시 구하는 방법을 사용했다. 각 특성에 가중치를 다르게 부여했을 때의 순위가 원래의 순위와 달라지는 것을 확인한다면 논문에서 제안한 사실에 정당성을 부여할 수 있고, 이를 호텔뿐만 아니라 여러 가지 제품과 서비스에 적용할 수 있을 것이다.

Python 웹 크롤링을 사용한 데이터 수집부터 다양한 라이브러리와 API를 사용해서 전처리, 극성 분석을 수행하고, 이를 통해 의미 있는 결과를 도출하기까지 모두가 새로운 시도의 연속이었다. 하지만 다양한 논문을 읽으며 오랜 기간 동안 많은 학자들의 오피니언 마이닝과 관련된 기술을 향상시키기 위해 노력한 것을 알게 되었다. 또한 이러한 노력이 이론적인 것에 그치지 않고 사람들의 삶에 가치 있게 적용되기 위해 노력해야한다는 것을 알았다.

■ 참고문헌

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

[2] 김유신, "주가지수 예측을 위한 뉴스 빅데이터 오피니언 마이닝 모형," 국민대학교 IT 전문대학원 박사학위 논문, 2012.

[3] 송종석, 이수원, "상품평 극성 분류를 위한 특징별 서술어 긍정/부정 사전 자동 구축," 정보과학회 논문지(소프트웨어 및 응용) 제38권 제3호, 2011.

[4] 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.

[5] Pang, B. and L. Lee, "Opinion mining and sentiment analysis," Foundations and Trends in Information Retrieval, Vol.2, 2018.

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

[7] 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.

[8] 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.

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

[10] Pavlopoulos, J., "Aspect Based Sentiment Analysis," Department of Informatics, Athens University of Economics and Business Ph.D Thesis, 2014.

[11] Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I. and Manandhar, S., "SemEval-2014 Task 4: Aspect Based Sentiment Analysis," Asian Journal of Computer Science & Information Technology, 2014.

[12] 한의환, 차형태, "Iterative Naive Bayes Classifier Design method for Effective Emotion Classification," 전자공학회논문지 제54권 제12호, 2017.