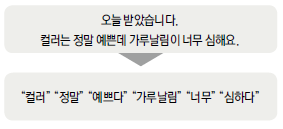
Ⅰ. 서 론

최근 글에 내재해 있는 사람들의 주관적 태도나 감성을 추출해 내는 분석 기법인 감성 분석에 대한 관심이 높아지고 있다. 감성 분석이란 텍스트에 나타난 주관적 요소인 긍정적 감정과 부정적 감정을 판별해 이를 수치나 도식, 등급으로 정량화하는 작업을 말한다. 감성 분석의 대표적 성공 사례로 초코바 스니커즈의 소비자 감성 변화에 따른 가격변동 마케팅, 국내 유명 화장품 브랜드 에뛰드하우스의 감성 분석 등을 꼽을 수 있다.[1] 하지만 물품을 구매할 때 리뷰에 의존하는 경우가 많기 때문에 이를 역이용해 리뷰 조작 업체들이 돈을 받고 허위 리뷰를 작성하는 경우도 존재하였다. 실제로 음식 중개 앱 ‘배달의민족’에서는 2개월간 2만5,000여건의 허위 리뷰를 검출하였고, 리뷰 조작 업체들을 수사기관에 고소하기도 했다.[2] 이에 우리는 감정 분석과 여러 지표를 활용하여 이러한 문제점을 해결하려한다.

Ⅱ. 본 론

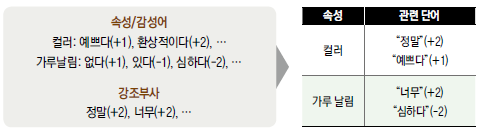
1. KoNLPy를 이용한 키워드 분석[3]

KoNLPy는 한국어 자연어 처리를 위한 Python 패키지이다. 웹 크롤링 과정에서 발생한 불필요한 단어와 의미를 알 수 없는 단어를 제거하였으며 숫자, 특수기호 등의 불용어를 제거하였다.

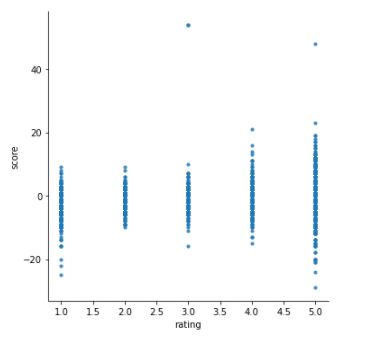


1. KNU 한국어 감성사전[4]

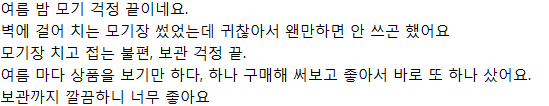
KNU 한국어 감성사전은 표준국어대사전을 구성하는 형용사, 부사, 동사, 명사의 모든 뜻풀이에 대한 긍정, 중립, 부정으로 분류하기 위해 Bi-LSTM 딥 러닝 모델 사용하였으며, 총 14,843개의 각 단어의 긍부정 판별은 (1) 매우 부정(-2) (2) 부정(-1) (3) 중립(0) (4) 긍정(+1) (5) 매우 긍정(+2) 등 리커트 척도(개인, 대상, 관념, 현상 등에 대한 개인의 태도나 성향의 강도를 측정하는 기법)를 이용하여 평가자들의 합의를 통해 선택된 방식이다.



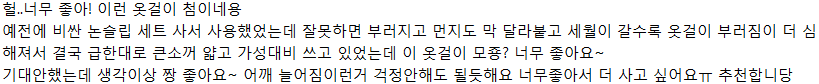
KoNLPY로 분리한 단어 중 KNU 한국어 감성사전과 일치하는 단어에 점수를 대입하여 각 리뷰별 감성 점수를 구할 수 있었다. 약 15만건의 리뷰를 분석해본 결과 별점이 높을수록 감성 점수가 높은 추세를 보였다. 전체 감성 점수의 평균값은 0에 수렴하였다.



하지만 높은 별점에서도 다수의 낮은 감성 점수도 검출되었으며, 이는 “부정적인 상황에서 해당 제품덕분에 좋아졌다”, ”다른 부정적인 리뷰에 걱정했지만 생각보다 좋은 제품이다.” 등의 리뷰로 판단된다. 이는 본 프로젝트에서 사용한 단어사전의 한계점으로, 각 분야별 감성분석에 필요한 맞춤형 감성사전 구축을 위한 추가적인 연구가 필요한 것으로 사료된다.





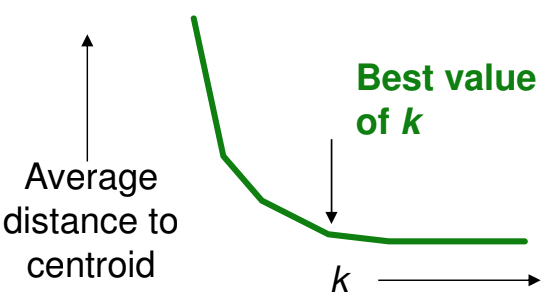
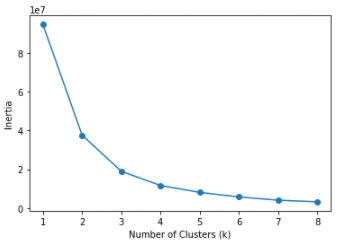


1. K-means 클러스터링

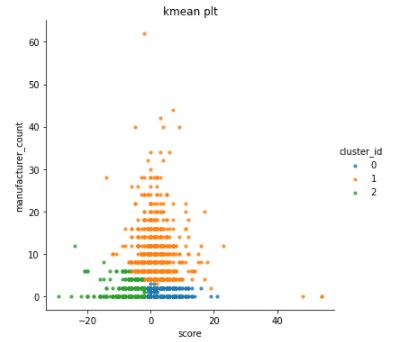
클러스터링은 유사성 따위의 개념을 바탕으로 데이터를 몇 개의 그룹으로 분류하는 기법을 칭한다. 이를 바탕으로 리뷰에서 추출한 특별한 지표들을 분석해보자 한다.

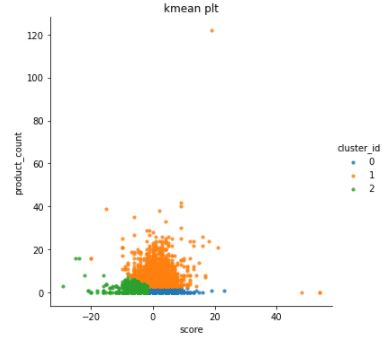
Spam 문자 메시지는 가능하면 의도한 바를 충분히 전달해야 하기 때문에 평균적으로 문장의 길이가 길고 정보를 구성하고 있는 단어 개수, 특수 문자 개수 등에 규칙성이 있음을 확인할 수 있었다.[4] 본 프로젝트에서는 해당 연구결과를 바탕으로 리뷰에 적용하여 언급된 제품명 수, 제조사명 수, 문장의 길이와 분석한 감성점수를 클러스터링하였다..

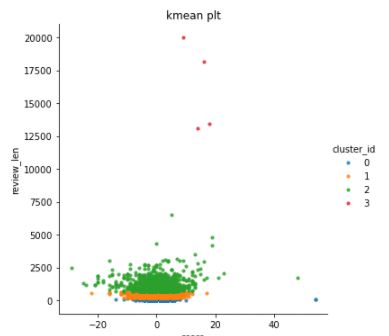
우선 k-means 클러스터링에서 최적의 k를 찾는다. 개수가 늘 때마다 평균이 급격히 감소하는데 적절한 k가 발견되면 매우 천천히 감소한다. 분석결과 대체로 최적의 k는 3인 것을 알 수 있었다.

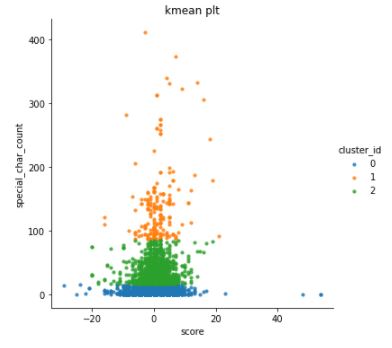
 

다음은 각 지표와 감성정수를 클러스터링해본 결과이다.

제조사명 언급횟수와 감성점수

제품명 언급횟수와 감성점수

리뷰길이와 감성점수

특수문자와 감성점수

Ⅲ. 결 론

분석 자료를 바탕으로 클러스터링 경계값을 구하여 그 이상에 속한 값은 가짜 리뷰로 판별하였다.

**감성점수 > 0, 특수문자 > 10, 제품명 언급횟수 > 3, 제조사명 언급횟수 > 3, 리뷰길이 > 300**

그 결과 238개의 리뷰를 판별할 수 있었다.

참고자료

[1] <https://dbr.donga.com/article/view/1202/article_no/8891/ac/magazine>

[2] <https://www.hankookilbo.com/News/Read/A2020112409210003295>

[3] ] KoNLPy: Korean NLP in Python, KoNLPy.org. Retreived at http://KoNLPy.org/en/latest/ on Aug. 15, 2019.

[4] Oh, Hayoung. (2016). 동시출현 단어분석 기반 스팸 문자 탐지 기법. 정보보호학회논문지, *26*(3), 693–700. https://doi.org/10.13089/JKIISC.2016.26.3.693