KIPS Trans. Softw. and Data Eng.
Vol.9, No.9 pp.259~266
ISSN: 2287-5905 (Print), ISSN: 2734-0503 (Online)

https://doi.org/10.3745/KTSDE.2020.9.9.259

A Technique for Product Effect Analysis Using Online Customer Reviews

Young Seo Lim† · So Yeong Lee†† · Ji Na Lee††† · Bo Kyung Ryu†† · Hyon Hee Kim††††

ABSTRACT

In this paper, we propose a novel scheme for product effect analysis, termed PEM, to find out the effectiveness of products used for improving the current condition, such as health supplements and cosmetics, by utilizing online customer reviews. The proposed technique preprocesses online customer reviews to remove advertisements automatically, constructs the word dictionary composed of symptoms, effects, increases, and decreases, and measures products' effects from online customer reviews. Using Naver Shopping Review datasets collected through crawling, we evaluated the performance of PEM compared to those of two methods using traditional sentiment dictionary and an RNN model, respectively. Our experimental results shows that the proposed technique outperforms the other two methods. In addition, by applying the proposed technique to the online customer reviews of atopic dermatitis and acne, effective treatments for them were found appeared on online social media. The proposed product effect analysis technique presented in this paper can be applied to various products and social media because it can score the effect of products from reviews of various media including blogs.

Keywords : Product Effectiveness Analysis, Effectiveness Measurement Algorithm, Online Customer Review

온라인 고객 리뷰를 활용한 제품 효과 분석 기법

임 영 서† · 이 소 영†† · 이 지 나††† · 류 보 경†† · 김 현 희††††

요 약

본 논문에서는 온라인 고객 리뷰를 활용하여 건강 보조제, 화장품 등 현재의 상태를 개선하기 위해 사용되는 제품을 대상으로 그 효과를 알아보기 위한 제품 효과 분석 기법을 제시하였다. 제안하는 제품 효과 분석 기법은 블로그 포스팅에 존재하는 광고를 자동 제거하고, 효과 분석을 위한 증상, 효과, 증가, 및 감소로 이루어진 단어 사전을 구축하며, 제안하는 알고리즘을 통해 제품의 효과를 측정한다. 제품 효과 분석 기법을 검증하기 위해 정답 레이블이 존재하는 네이버 쇼핑 리뷰 데이터셋을 대상으로 성능평가를 실시하였으며, 전통적인 긍부정 사전과 RNN 모델과 성능을 비교하였다. 실험 결과, 본 논문에서 제안하는 효과 분석 기법이 다른 두가지 방법보다 정확도가 뛰어남을 보여주었다. 또한, 아토피 피부염, 여드름 치료제에 제안하는 기법을 적용하여 소셜 미디어에 나타난 효과적인 치료법을 소개하였다. 본 논문에서 제시한 알고리즘은 블로그를 포함한 여러 매체의 리뷰로부터 제품의 효과를 점수화할 수 있으므로 다양한 제품군과 소셜 미디어에 적용될 수 있을 것으로 보인다.

키워드 : 제품 효과 분석, 효과 측정 알고리즘, 온라인 고객 리뷰

1. 서 론

최근 온라인 상품 리뷰를 분석하여 마케팅에 활용하거나 고객 서비스에 적용하는 등 소셜 미디어에 대한 감성 분석이 활발히 연구 및 활용되고 있다[1]. 대부분의 온라인 고객 리뷰에 대한 감성 분석 연구들은 다양한 분야에서 자연어 처리 기술을 적용하여 좋은 성과를 보이고 있다[2, 3]. 하지만 건강보조제나 화장품 등 효과를 볼 수 있는 제품을 서술하는 리뷰는 사용되는 형용사가 증상과 함께 사용되었는지, 효과와 함께 사용되었는지에 따라 그 의미가 달라져 기존의 감성 분석 기법을 적용하는 데 어려움이 있다. 또한, 한 개의 후기 내에 상품에 대한 다양한 측면의 평가가 서술되어 있으므로 문장 단위뿐만 아니라 한 문장 내의 관점 단위의 감성 분석이 필요하다.

본 논문에서는 기존의 감성 분석과는 달리 상품의 후기에 담긴 효과를 정확히 판단하기 위해 증상, 효과, 증가, 감소로 구성된 단어 사전을 구축하고 이를 기반으로 효과 점수를 측

※ 이 논문은 2019년도 동덕여자대학교 교내연구비 지원으로 연구되었음.
※ 이 논문은 2019년도 한국정보처리학회 추계학술발표대회에서 '소셜미디어를 활용한 아토피 치료법 효과 분석 모델'의 제목으로 발표된 논문을 확장한 것임.
† 준 회 원 : 동덕여자대학교 정보통계학과 학사과정
†† 비 회 원 : 동덕여자대학교 정보통계학과 학사과정
††† 비 회 원 : 동덕여자대학교 문헌정보학과 학사과정
†††† 정 회 원 : 동덕여자대학교 정보통계학과 부교수
Manuscript Received : April 22, 2020
First Revision : June 17, 2020
Second Revision : July 22, 2020
Accepted : July 28, 2020
* Corresponding Author : Hyon Hee Kim(heekim@dongduk.ac.kr)

정하는 알고리즘을 개발하였다. 또한, 광고가 섞여 있는 경우에는 분석 결과에 대한 신뢰도가 떨어지게 되므로 이를 제거하기 위한 광고 제거 작업도 수행하였다.

먼저, 효과 측정을 위해 기존의 긍정 및 부정어를 정의하는 기존의 한국어 감성 사전[4]과는 달리 증상, 효과, 증가 및 감소로 구성된 단어 사전을 구축하였다. 제품의 효과에 대한 리뷰들은 긍정 및 부정을 나타내는 단어가 증상과 함께 사용될 경우 및 효과와 함께 사용될 경우 그 의미가 달라지므로 4개의 범주로 확장하여 단어 사전을 구축하였다. 사전에 등록될 단어는 TF-IDF를 활용하여 핵심이 되는 단어를 찾아내고, 핵심 단어로부터 Word2Vec을 활용하여 기존 사전의 단어들과 유사한 단어들을 찾아 사전 확장에 활용했다.

상품 리뷰를 분석할 때 다수의 글이 광고성 글들이므로 이러한 광고글을 찾아내 제거한 뒤 분석하는 것이 필수적이다. 따라서 본 논문에서는 html 태그, 정규 표현식 등을 이용하여 광고를 자동으로 삭제할 수 있는 광고 제거 알고리즘을 개발하여 분석의 정확도를 낮추는 광고를 사전에 제거하였다.

마지막으로 구축된 사전을 기반으로 각 리뷰의 효과 점수를 측정하는 제품 효과 측정 알고리즘(Product Effectiveness Measurement Algorithm, 이하 PEM)을 제안하였다. 본 논문에서 제안하는 PEM 알고리즘은 앞에서 언급한 네 가지 단어특성을 고려하여 문장에서 사전이 조합되는 경우의 수에 따른 점수를 계산하고, 이를 본문 전체의 점수로 확장하는 알고리즘이다

제안하는 제품 효과 분석 기법의 성능평가를 위해 네이버 쇼핑의 댓글 데이터를 활용하여 기존의 KNU 사전과 RNN을 적용한 결과를 비교하였다. 그 결과, 건강 보조 식품으로 사용하는 제품에 대해서는 타 분석 기법들과 유사한 성능을 보였으나, 특정 제품의 효능을 기대하고 구매하는 기능성 화장품에 대해서는 PEM 알고리즘이 더 나은 결과를 보였다.

성능평가의 결과를 토대로 PEM 알고리즘을 실제 데이터에 적용해보기 위해 아토피, 여드름과 같이 증상 치료를 위해 처방받은 의약품 외에도 화장품, 영양제와 같은 다양한 수단을 보조적으로 사용하므로 분석할 수 있는 제품군이 다양하여 적합하다고 판단하였다. API를 활용해 선정된 키워드에 따라 네이버 블로그의 '아토피 치료', '여드름 치료'를 주제로 분석할 데이터를 수집하였다. 이후, 전처리 과정을 거친 블로그 데이터를 본 연구에서 제안한 감성 분석모델에 적용하였다.

본 논문의 공헌은 적용 사례로 든 아토피 치료, 여드름 치료와 같이 효과를 알고자 하는 여러 제품군의 효과를 분석하는 데에 있어 효율적인 감성 분석모델을 제안한 것이다. 이를 위해 네이버 블로그 포스팅 중 분석 대상과 관련된 정보를 수집하였고, 광고를 제거함으로써 신뢰도를 높일 수 있었다. 또한, PEM 알고리즘을 아토피 치료법과 여드름 치료법에 적용하여 효과가 뛰어난 치료법을 제시하였다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구, 3장에서는 본 논문에서 제시한 효과 측정 방법을 자세히 설명한다. 4장에서는 PEM 알고리즘과 기존의 감성 분석 기법과의 성능평가 결과를 제시하였다. 5장에서는 실제 데이터 적용 사례와 그 결과에 관해 설명하며, 마지막으로 6장에서는 결론을 제시한다.

2. 관련 연구

국내에서 이루어진 감성 분석 연구는 주로 트위터, 페이스북 등의 소셜 미디어 데이터를 대상으로 감성 분석이 이루어지고 있으며 이외에도 블로그, 웹 카페 댓글 등이 활용되고 있다[5]. 특히, 트위터 텍스트 자료를 활용한 텍스트 감정 분류 연구[6], 네이버 블로그 및 카페 텍스트 자료를 토대로 프로농구 발전 방안을 제시한 연구[7], 그리고 네이버 영화 후기 문장을 수집해 감성 분석을 시도한 연구[8] 등 소셜 미디어를 통해 감성 분석한 연구라는 점에서 본 연구와 밀접한 관련이 있다.

[6]의 연구는 트위터 텍스트 자료에서 검색 키워드를 추출한 후 데이터 필터링을 거쳐 감정 단어를 추출하고 네 가지 회귀분석 방법을 이용하여 분류하고 이를 바탕으로 감성 분석을 진행하였다. [6]은 감정 단어의 정도에 따라 긍정이라는 감정을 기쁨, 분노, 슬픔, 즐거움의 네 가지의 감정으로 분석하여 연구하였다.

[7]의 연구는 블로그 및 카페 포스팅을 수집하여 키워드 빈도 분석과 의미연결망 분석을 기반으로 프로농구 발전 방안을 제시하였다. 이를 통해 긍·부정 감성 요인의 개발 및 개선 방안을 알 수 있다. [7]의 연구에서는 감성 요인에 대해 의미연결망 분석 결과를 제시하였는데 단어 자체만 고려한 것이 아니라 단어의 문맥을 고려했다는 특징이 있다.

본 연구에서는 [6]과 [7]의 연구 방법을 빌려 유사한 방식으로 효과의 정도를 분석하기 위해 효과, 증가, 증상, 감소의 네 가지 특성을 가진 사전을 구성하였다. 더불어 이를 구성하기 위해 5가지의 형태소를 사용함으로써 단어의 문맥을 고려하여 효과를 긍정, 부정의 두 가지 감정으로 나타내고자 하였다.

마지막으로 [8]은 네이버 영화 후기 문장을 수집해서 감성 분석을 수행한 연구이다. 기존 감성 분석과는 다르게 감성 사전을 구축하지 않고 말뭉치로만 감성을 예측하고 음소와 형태소, 어절 단위 모형을 구축해 입력 단위에 따른 모형성능을 비교하였다. 하지만 이로 인해 긍·부정을 나누는 기준이 모호하다고 판단하여 본 연구에서는 이 점을 보완하여 정확성을 위해 한 문장 단위로 효과를 분석하는 작업이 필요할 것으로 보인다.

3. 제품 효과 분석 기법

본 논문에서 제안하는 효과 분석 기법은 자료 수집 시 포함된 광고를 제거하는 광고 글 제거 전처리, 증상, 효과, 증가, 감소의 표현으로 구성된 사전 구축, 그리고 효율적으로 각 문장의 효과 점수를 계산하는 PEM 알고리즘으로 구성되어 있다. 3.1절에서는 광고 글 제거를 위한 전처리 과정을 설

명하고, 3.2절에서는 사전 구축 방법을 자세히 서술한다. 마지막으로 3.3절에서 PEM 알고리즘을 제시한다.

3.1 광고 제거를 위한 전처리

광고를 제거하는 알고리즘은 다음과 같이 구현하였다. 먼저 브랜드(업체)에서 운영하는 홍보용 블로그는 특성상 블로그 이름에 브랜드(업체)와 관련된 단어가 포함되어 있다. 그러므로 이와 관련된 단어를 수집하여 리스트를 만든 뒤, 블로그 이름의 html 태그에 해당 단어가 포함되면 자동으로 수집 대상에서 제외하였다[2]. 이와 달리 개인이 협찬을 받아 블로그에서 제품을 광고하는 경우[9], 공정거래위원회의 지침에 따라 '본 게시물은 **에서 제품을 무(유)상제공 받아 작성된 글입니다.'와 같은 문구를 삽입해야 하며, 블로그 운영자가 상품을 홍보하며 구매를 종용하는 경우에는 구매는 문의를 달라는 문구 또는 구매링크가 삽입되어 있다. 이러한 포스팅은 정규 표현식을 활용하여 제거하였으며, 이 같은 2가지 방식으로 광고를 수집 대상에서 제외하도록 하였다. 광고 포스팅의 경우 비슷한 내용의 글을 반복한 포스팅을 다수 게재하는데 위와 같은 방식을 활용하면 모두 제거가 가능하다.

3.2 사전 구축

본 연구에서는 효과의 긍정, 부정을 가리기 위해 기존 감성 사전과 달리 증상(symptom), 효과(effect), 증가(up), 감소(down)로 구성된 새로운 사전을 구축하였으며 [10]에서 정의된 사전을 확장하였다. 사전에 들어가는 형태소의 종류는 명사, 동사, 형용사, 관형사, 부사이다. 위 형태소 중 명사, 동사, 일부 형용사는 증상이나 효과를 설명하는 형태소이며, 문장을 구성하는 주성분에 해당한다. 일부 형용사, 관형사, 부사는 증가, 감소의 정도를 의미하며 주성분을 수식하는 부속성분에 해당한다. 이러한 특징을 이용하여 5가지 형태소를 사용함으로써 다양한 문장 구조를 고려할 수 있도록 하였다. 각 형태소는 한국어 형태소 분석기인 Okt 패키지를 사용하여 정규화시킨 뒤, 이들을 활용하여 사전을 구축하였다.

사전 구성 전, 본 연구는 기존 감성 분석보다 더 많은 형태소를 고려하며, 모든 형태소가 효과와 정도의 의미를 지니고 있지 않으므로 보다 효율적으로 사전을 구성하기 위해 Term Frequency-Inverse Document Frequency (TF-IDF)를 활용하였다. 본 연구에서 분석 대상으로 하는 특정 키워드의 효과에 대한 포스팅들은 작성되는 텍스트의 길이가 각기 다르고 그 안에 다루는 주제가 다양하므로 형태소의 빈도를 나열했을 때 효과와 관련된 단어는 빈도가 낮은 쪽에 속한다. 그러므로 단순 빈도로는 사전을 구성하기에 비효율적일 것으로 판단하여, 특정 단어가 해당 문서 내에 등장한 빈도인 TF 값과 그 단어가 등장한 문서 빈도의 역수 값인 IDF 값을 곱한 값인 TF-IDF [11]을 활용하였다. TF-IDF 값을 이용해 단어들을 정렬했을 때 단순 빈도를 사용하는 것보다 수월하게 핵심어를 찾아냄으로써 효율적으로 사전을 구성할 수 있었다.

위의 방법을 이용하기 위해 Counter 패키지를 사용하여 출현빈도가 2회 이상인 단어들을 추출한 뒤, Term Document Matrix(TDM)를 생성하였다. 이후, TDM을 TF-IDF로 변환하였고 이를 가중치의 내림차순으로 정렬하였다. 이렇게 정렬된 단어는 점수와 함께 네 가지 특성으로 구성된 사전을 구축하는 데 활용되었다. 부정적 의미를 담고 있는 증상, 감소 사전은 음수인 -1의 점수를 부여하였으며, 긍정적 의미를 담고 있는 효과, 증가 사전은 양수인 +1을 점수로 사용하였다. 제작한 사전의 예시는 아래 Table 1과 같다.

Table 1. Examples of the Weight Dictionary

Symptom	Effect	Up	Down
어지러움	호전	너무	불편하다
붉어짐	개선	늘어나다	없어지다
가려움	환하다	갖다	적다
쑤시다	효과적	많이	실패하다
열감	면역력	떠오르다	아쉽다
구토	반응	밝다	그치다
매스꺼움	진정	생기다	가라앉다
경련	좋아지다	높다	누그러지다
트러블	보습	나타나다	끊다
울긋불긋	편안하다	생기다	처지다

TF-IDF를 활용하는 과정에서 중복적인 단어가 자동으로 제거되었고, 중요도가 낮은 불필요한 단어들을 한눈에 알아볼 수 있어 사전 제작의 효율성을 높일 수 있었다.

이후 PEM 알고리즘의 성능을 향상하고자 사전의 확장을 진행하였다. Gensim의 Word2Vec 패키지를 활용하여 수집한 데이터에서 기존 사전에 포함된 단어들과 유사한 단어들을 추출하였다. 사전이 효과, 증상, 증가, 감소 등 데이터 내에 중심이 되는 단어들로 구성이 되어있어 주변 단어들로 중심단어를 예측하는 Continuous Bag of Words (CBOW) 방식을 활용하였다. 수집한 데이터에서 각 사전의 단어들과 유사도가 높은 단어들을 찾아 점수와 함께 각 사전에 추가하였다.

3.3 효과 측정 알고리즘(Product Effectiveness Measurement)

제3.2절에서 설명한 바와 같이 제작된 사전을 활용하여 효과를 수치화하는 PEM 알고리즘은 효과, 증상, 증가, 감소 사전 간의 관계를 조합하여 형태소, 문장, 포스팅 순으로 계산되는 알고리즘이다. 해당 알고리즘은 하나의 문장을 기준으로 점수를 계산하는 아래 Table 2의 line 5에서 line 21의 효과 점수 측정과 이를 반복하여 하나의 포스팅(p), 전체 포스팅(f)으로 점수를 확장하는 식으로 구성된다.

먼저 문장 단위의 효과 점수 측정을 위해 포스팅의 각 문장은 효율적인 분석을 위해 한국어 형태소 분석기인 Okt를 이용하여 품사별로 나누어져 sentence라는 리스트에 저장된다. 나누어진 단어에 해당하는 사전의 이름(kinds)과 각 단어의 고유점수(weight)가 저장된 형태로 전처리 과정을 거친 뒤, 알고리즘에 입력된다.

예를 들어 '가려움이 늘어났어요.'라는 문장은 증상 사전에 해당하는 '가려움'과 증가 사전에 해당하는 '늘어나다'와 같이 정규화된 단어가 추출된다. 이들은 각 단어의 고유점수와 함께 [['Symptom', -1], ['Increase', +1]]의 형태로 재구성되는 것이다. 재구성된 문장은 아래 Table 2의 알고리즘 내에서 line 5의 sentence 변수로 호출된다. 이를 기준으로 증상, 효과 사전에 해당하는 단어 뒤에 증가, 감소 사전에 포함된 단어가 연속될 때에 해당하는 조합에 따른 계산하는 과정을 거쳐 score에 저장된다.

최종적으로 위의 예시인 '가려움이 늘어났어요.'라는 문장은 효과 점수 측정 알고리즘을 통해 효과 점수 -1, increase 점수 +1을 획득하게 되며, 최종적으로 이를 곱한 점수인 (-1) * (+1) = -1이 score 점수로 반환된다.

Table 2. Product Effectiveness Measurement Algorithm

```
Algorithm PEM

Set s_score, s_num is 0
Set p_score, f_score is []
For posting, full_text do
    For sentence, posting do
        For i = 0, len(sentence) do
            Set p_pos, c_pos, score, effect, increase is 0
            If (kinds = "Symptom" or "Effect")
                Set c_pos is 1
                If (p_pos = 0)
                    Set score is effect * increase
                Else if (p_pos = 1)
                    Set score is effect + weight
                Set p_pos is 1
            Else if (kinds = "Increase" or "Decrease")
                Set c_pos is 0
                If (p_pos = 0 and weight*increase < 0)
                    Set score is increase + weight
                Else if (p_pos = 0 and weight*increase > 0)
                    Set score is increase * weight
                Set p_pos = 0
            Set effect, increase is 0
        Set s_score is s_score + score
        If (score is not 0)
            Set s_num is s_num+1, score is 0
    Set p_score to p_score.append(s_score/s_num)
    Set s_score to 0
Set f_score to f_score.append(p_score)
Return f_score
End
```

또한, 중첩 if 문과 과거 형태소의 사전종류를 알기 위한 p_pos, 현재 형태소의 사전종류를 알기 위한 c_pos 변수를 이용하여 증상(효과) * 증가(감소)조합 이외의 조합에 대해서도 계산될 수 있도록 하였다. 다양한 형태소를 활용함으로써 발생할 수 있는 같은 사전의 반복, 증상 * 효과와 증가 * 감소와 같이 서로 대조되는 조합까지 고려함으로써 더욱 다양한 문장 구조에 대해 계산할 수 있다. 예를 들어 '힘들지 않다.'라는 문장은 형태소 분석기를 통해 증가 사전에 해당하는 '힘들다'와 감소 사전에 해당하는 '않다'로 나누어진다. 단어들이 각각 양수, 음수인 점수를 가지므로 더하면 점수가 상쇄하여 0에 수렴하는 결과를 얻는다. 이와 다르게 '없지 않다.', '그치지 않다,'와 같은 이중 부정문은 긍정의 의미를 띄는데 본 연구에서는 '없다', '그치다', '않다' 단어들이 모두 감소 사전에 해당한다. 그러므로 이들이 가진 음수 고유점수들을 곱하면 양수 점수를 가지게 되므로 긍정의 의미로 분석된다.

이처럼 line 5에서 line 21에 걸친 효과 점수 측정은 각 문장이 4가지 사전의 조합에 따라 계산됨으로써 한 문장 안에서도 여러 번의 계산이 이뤄져 더욱 다양한 효과 점수를 얻을 수 있다.

효과 점수 측정에서 각 문장의 분석 결과인 score는 PEM의 s_score(문장 점수)에 누계된다. 그렇게 하나의 포스팅에 대한 분석이 끝나면 포스팅 길이에 따라 문장 점수의 합이 크게 달라짐을 고려하여 포스팅 점수들이 같은 범주 내에서 결과를 낼 수 있도록 s_score를 해당 포스팅의 유효한 문장의 수인 s_num으로 나누어 변환된 값을 포스팅 점수로 사용한다. 해당 계산이 끝나면 p_score(포스팅 점수 리스트)에 추가한다.

위와 같은 방식으로 키워드 내에 모든 포스팅의 p_score가 계산되면 이를 f_score(키워드 점수 리스트)에 추가한다. 해당 과정들을 반복하면 최종적으로 키워드별 점수 리스트인 f_score를 얻을 수 있다. 본 논문의 PEM은 https://github.com/lim1014/Effect-Analysis-Algorithm 에서 다운로드하여 실행이 가능하다.

4. 성능평가

기존의 감성 분석은 트위터 또는 댓글과 같은 성격의 짧은 텍스트의 긍정, 부정을 판별하기엔 효과적이다. 그러나 블로그 데이터와 같이 하나의 포스팅 속에 다양한 주제의 글을 포함하는 경우, 글의 길이도 길기에 보편적인 감성 분석방법이 적합하지 않다. 본 논문에서는 길이가 긴 텍스트의 감성 분석과 동시에 다양한 주제 중 '효과'라는 속성에 대해 특화된 주제에 대해 성능평가를 실시하였다.

4.1 성능평가 데이터

성능평가를 위한 데이터로 길이가 다양하면서 동시에 사용자가 직접 매긴 점수가 존재하는 소셜커머스의 제품 리뷰 데이터를 활용하였다. 소셜커머스 매체로는 네이버의 쇼핑몰을 사용하였는데, 이는 네이버가 2019년 1/4분기 국내 소셜 미디어 상위 10위에 관련 매체 3개가 자리매김하였으며, 상품 후기 작성 시 작성된 리뷰의 길이, 형태에 따라 즉시 현금화가 가능한 포인트를 차등지급함으로써 정성 들인 리뷰가 타 소셜커머스보다 다수 존재한다고 성능평가를 위한 데이터로

적합하다고 판단하였기 때문이다.

분석 대상이 되는 제품으로는 아토피성 피부와 같이 손상된 피부를 개선하기 위해 사용하는 기능성 화장품(Functional Cosmetics), 건강보조식품으로 제일 많이 섭취하는 멀티비타민(Multi Vitamin), 유산균(Lactobacillus)을 이용하였다. 각각의 카테고리에서 리뷰 수가 많은 제품에 대해 리뷰 데이터를 크롤링하였다. 크롤링한 리뷰 데이터의 개수는 기능성 기초화장품에 대해서는 1,523개, 유산균은 3,090개, 멀티비타민은 2,617개이다.

Selenium 패키지를 이용하여 크롤링한 데이터는 사용자가 매긴 점수(최소 1점~최대 5점)가 4점, 5점에 해당하는 리뷰 수가 대부분을 차지하는 쏠린 정규분포(Skew Norm)를 띄고 있으며, 일반적으로 보통의 범주에 해당하는 3점을 받은 리뷰의 대체적인 의견이 그저 그렇다, 효과가 있는지 모르겠다 등의 의견이었다. 그리하여 주로 추천, 괜찮다, 적당하다.의 의견에 해당하는 4점, 5점 리뷰에 대해 긍정(Positive)으로, 이와 반대로 그저 그랬다, 좋은지 모르겠다.의 의견에 해당하는 1점, 2점, 3점 리뷰에 대해서는 부정(Negative)으로 분류하였으며 그 개수는 아래 Table 3과 같다.

Table 3. Crawling Result

Functional Cosmetics	Positive	1,355
	Negative	168
Lactobacillus	Positive	1,700
	Negative	691
Multi Vitamin	Positive	1,600
	Negative	1,017

4.2 성능평가 결과

성능평가를 위한 실험은 파이썬 언어를 사용하여 윈도우 10 환경에서 실행되었다. 성능평가를 시행하기 위해 한국어 감성 분석 사전인 KNU 사전 기반 감성 분석과 텍스트 분류에 유용한 RNN을 선정하여 비교하였다. 3개의 제품 리뷰 데이터의 분석을 위해 파이썬의 Tensorflow 패키지를 활용하여 드롭아웃 0.2, 이진 분류에 해당하므로 출력층의 활성화 함수로 Sigmoid, 최적화 함수 Rmsprop, 손실함수 Binary_Crossentropy를 부여하여 RNN 모델을 제작하였다. 전체 데이터의 70%를 훈련데이터로 동시에 학습시켜 정확도 64.79%와 손실 함숫값 0.5411을 기록하였다. 성능 비교를 위해 각각의 분석모델에 3가지 제품의 남은 30%의 시험데이터를 적용하였으며, KNU 한국어 감성 사전과 PEM 알고리즘의 경우 분석 결과가 점수로 제시되므로 레이블을 같이 부여하기 위해 0점 이상인 리뷰에 대해서는 긍정, 0점 미만인 리뷰에 대해서는 부정을 부여하였다.

3가지 제품의 각각의 분석 기법이 예측한 결과에 대한 정확도를 그림으로 나타낸 결과는 아래 Fig. 1과 같다.

Fig. 1의 막대 그래프는 왼쪽부터 순서대로 기능성 화장품,

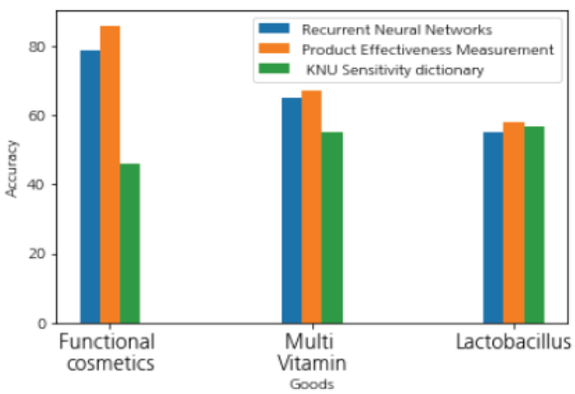

Fig. 1. Accuracy Graph for Performance Tests

멀티비타민, 유산균을 가리키며, 각각의 정확도는 기능성 화장품의 경우 RNN에서 79%, PEM은 86%, KNU에서 46%를 나타내었으며, 멀티비타민에 대해서는 RNN 65%, PEM 67%, KNU 55%, 유산균에 대해서는 RNN 55%, PEM 58%, KUN 57%를 나타내었다.

그림에서 보이는 바와 같이 유산균, 멀티비타민에 대해서는 근소한 차이로 PEM 알고리즘 적용 결과가 조금 더 실제 리뷰의 반응과 유사하게 예측이 되었음을 알 수 있다. 하지만 기능성 화장품에 대해서는 앞의 두 제품에 비해 세 분석방법 간의 차이가 확연하게 드러나는 것을 볼 수 있다.

기능성 화장품의 경우, 제품이 말하는 효과와 기능에 대해 기대를 하고 구매로 이어지는 경우가 대다수이다. 그러므로 리뷰에서도 단순하게 추천, 비추천을 말하는 후기보다는 제품의 효과, 기능에 대해 상세히 언급한 후기가 대부분이었기 때문에 텍스트에 나타난 효과의 감성을 계산하는 PEM 알고리즘이 이를 더 잘 파악한 결과라고 볼 수 있다. 하지만 유산균, 멀티비타민은 건강보조식품으로 체질 개선이나 건강을 위해 일부러 먹는 것이기 때문에 효과에 대한 후기가 비교적 적었다. 하지만 감성 분석의 최신 기법인 RNN, 본 논문과 같이 사전을 활용하여 감성을 분석하는 KNU 한국어 감성 사전보다는 높은 정확도를 나타낸 것을 보아, 효과를 분석하는 데에서는 본 논문의 알고리즘이 더 유용하다 판단된다.

4.3 연구 결과 적용

제5장에서는 실제 데이터를 PEM 알고리즘에 적용하여 분석해보고자 한다. 제4장에서 실시한 성능 평가의 결과를 보면 알 수 있듯이 PEM 알고리즘은 텍스트의 효과를 분석하는데 적합한 알고리즘이다. 이를 알고리즘의 구축 배경이었던 네이버 블로그에 적용해보고, 더불어 성능 평가와 달리 제품이 아닌 대상에 대한 후기에서도 효과를 분석해보고자 하였다.

'아토피'라는 질병은 장기적인 치료가 필요한 경우가 많으며, 일반적으로 증상 치료를 위해 병원에서 의약품을 처방받아 사용한다. 하지만 장기적인 치료에 소수의 의약품에서 우려되는 부작용을 피하려고 의약품 외에도 보조적으로 화장품, 영양제, 민간요법을 이용하는 경우가 많다. 그러므로 '아

토피 치료'는 그 효과를 분석할 수 있는 키워드가 다양하여 본 연구에서 제안한 모델에 적합하다고 판단하였다[12].

그리고 실제 데이터 적용 예시로 '아토피' 외에 사춘기에 피지분비 증가로 나타나는 비염증성 또는 염증성 피부 발진인 '여드름 치료'[13]에 대해 같이 분석하였다. 또한, 여드름은 아토피와 유사하게 상황에 따라 장기적인 치료가 요구되기도 하는 대다수 사람이 흔하게 겪는 피부 질환으로 의약품 이외에 보조적 요소를 사용하여 증상을 완화하려고 하는 경우가 많다는 것이다.

4.4 아토피 치료법 적용

'아토피 치료'와 관련된 새로운 키워드와 기존에 자주 언급되는 키워드를 선정하기 위해 파이썬의 BeautifulSoup 라이브러리와 request 패키지를 활용하여 뉴스 크롤링을 진행하였다. 네이버 뉴스에서 '아토피 치료'를 검색한 뒤, 이에 해당하는 기사 51,648개의 본문 내용 전체를 수집하였다(2020.03.29. 기준). 형태소 분석기 Okt를 활용해 수집된 데이터에서 한국어 명사를 추출하고, Counter 패키지를 이용하여 아토피 치료성분과 치료법에 관련된 상위 20개의 단어를 선별하였다. 이들 중 충분한 데이터가 존재하는 최종적으로 '락토바실러스', '항히스타민', '플라즈마' 등 6개의 키워드가 선정되었다. 네이버 API와 파이썬 BeautifulSoup 라이브러리를 이용하여 각각의 키워드를 기준으로 481~67,124개의 글에 대해 블로그 이름과 본문을 수집하였다(2020.03.29. 기준).

데이터 수집 결과 확인된 다수의 광고 포스팅을 제거하기 위해 본 논문의 제3장에서 언급한 광고 제거 알고리즘을 사용하였다. 광고 제거 알고리즘을 적용한 결과 수집한 데이터에서 광고 포스팅이 제거되었다. 키워드별로 살펴보면 감마리놀렌산 77.2%, 듀피젠트 61.9%, 락토바실러스 87.2%, 스테로이드 78.2%, 플라즈마 79.6%, 항히스타민제 75.7%에 해당하는 포스팅이 광고로 분류되어 제거되었으며 이를 통해 더 객관적인 데이터를 얻을 수 있게 되었다. 광고를 제거한 데이터는 분석이 원활하게 이루어질 수 있게 정규식을 활용하여 'ㅋㅋㅋ', 'ㅎㅎㅎ' 과같이 반복되는 단일문자와 '^^', '~'과 같은 문장기호, 이모티콘, 과도한 공백 등 불필요한 문자를 제거하였다.

전처리가 완료된 데이터에 4가지의 사전을 활용하여 PEM 알고리즘을 적용하였다. 그 결과, 각각의 키워드의 이상치를 포함한 포스팅의 효과 점수 분포는 다음 Fig. 2와 같다.

Fig. 2의 Box plot을 살펴보면 전체적으로 10개 미만의 이상치가 존재하며, 편차가 크지 않다. 이는 각 포스팅의 길이에 따른 점수의 범위가 크게 달라지지 않도록 포스팅 문장의 합을 해당 포스팅의 유효한 문장의 수로 나누어 변환된 값을 포스팅 점수로 활용한 결과이다.

이렇게 실제 블로그 후기 중 약 76.6%의 광고성 글을 제거한 후기들을 PEM을 통해 분석한 결과는 다음과 같다. Fig. 2에서 보이는 것과 같이 다른 키워드에 비해 0점보다 낮은 점수가 많은 일반적인 치료제로 알려진 스테로이드와 항히스타민에 대해서는 장기간 사용에 따라 발생하는 부작용과 내성으로 인해 부정적인 글들이 많이 나타났으며[14], 반대로 비교적 높은 점수가 많이 분포한 달맞이꽃 기름의 성분인 감마리놀렌산, 물리적 치료기기의 물질인 플라즈마, 유산균의 성분인 락토바실러스는 이른 시일 안에 효과는 볼 수 없지만 염려되는 부작용이 적어 긍정적인 후기의 비율이 더 높은 것으로 나타났다. 하지만 '듀피젠트'의 경우 점수의 분포가 다른 키워드보다 훨씬 넓게 나타났음을 알 수 있다. 해당 키워드는 중증 아토피 환자에게 사용되는 약물로 고가의 의약품이며, 국내에 출시된 지 약 1년 반 정도가 되었다. 그러므로 이를 사용한 사람이 다른 치료법에 비해 많지 않은 만큼 효과에 대해 비교적 극단적인 부분이 있어 다른 치료법과 비교해 분포가 넓게 나타난 것으로 보인다.

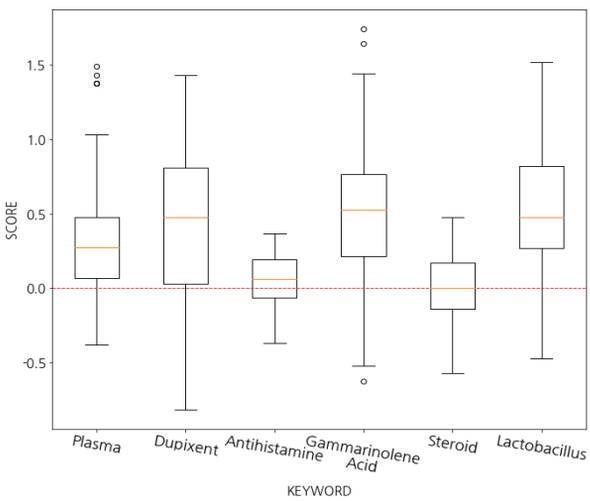

Fig. 2. Boxplot of Results of the PEM Algorithm

4.5 여드름 치료제 적용

'여드름 치료'의 경우에도 이전과 같은 방식으로 6,939개의 기사를 크롤링하여 키워드를 선정하였다. 선정된 키워드의 블로그 데이터 수집 및 전처리 과정과 PEM 적용을 거쳐 빈도수가 높은 단어인 '살리실산', '이부프로펜피코놀' 등 4개의 키워드를 선별하여 296~4728개의 블로그 본문 데이터를 수집하였다.(2020.07.14. 기준) 광고 제거 알고리즘을 사용하여 수집한 데이터에서 키워드별로 살리실산 72.2%, 이부프로펜피코놀 89.9%, 스테로이드 96%, 플라즈마 82.4%에 해당하는 포스팅이 광고로 분류되어 제거되었다. 광고 제거와 더불어 불필요한 문자의 전처리가 완료된 데이터에 4가지의 사전을 활용하여 PEM을 적용하였다. 그 결과, 각각의 키워드의 이상치를 포함한 포스팅의 효과 점수 분포는 다음 Fig. 3과 같다.

'여드름 치료'의 경우에도 이전과 같은 방식으로 자료 수집 및 전처리 과정과 PEM 적용을 거쳐 빈도수가 높은 단어인 '살리실산', '이부프로펜피코놀' 등 4개의 키워드를 선별하여 296~4728개의 블로그 데이터를 수집하였다. 광고 제거 알

고리즘을 사용하여 수집한 데이터에서 키워드별로 살리실산 72.2%, 이부프로펜피코놀 89.9%, 스테로이드 96%, 플라즈마 82.4%에 해당하는 포스팅이 광고로 분류되어 제거되었다. 광고 제거와 더불어 불필요한 문자의 전처리가 완료된 데이터에 4가지의 사전을 활용하여 PEM을 적용하였다. 그 결과, 각각의 키워드의 이상치를 포함한 포스팅의 효과 점수 분포는 다음 Fig. 3과 같다.

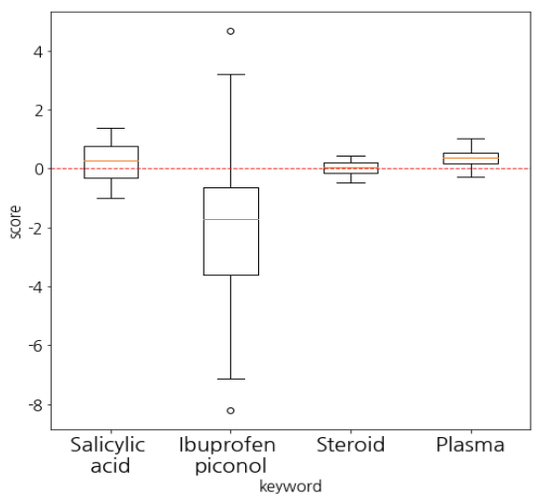

Fig. 3. Boxplot of Results of the PEM Algorithm

Fig. 3에서 부정적인 후기가 많은 것으로 보이는 이부프로펜피코놀과 살리실산의 경우 항균효과, 항염증 작용을 하여 여드름을 포함한 피부 질환에 효과가 있다고 한다. 하지만 실질적인 피부 질환 치료에 대해서는 스테로이드성 성분보다 개선 효과가 그렇게 크지 않으며, 해당 성분이 들어간 연고 등을 사용할 때에 바른 부위가 자극을 받아 붉게 달아오르거나 건조함 등의 부작용이 발생할 수 있어 이에 대한 부정적 후기가 영향을 미친 것을 보인다.

반면, 스테로이드의 경우 아토피 치료의 후기와는 다른 반응을 보이는 것을 확인할 수 있다. 스테로이드는 아토피 치료에서 일반적인 치료제로 쓰이는 것과 달리 여드름 치료에서는 심한 염증으로 인해 피부가 딱딱해지는 등 상태가 중증 정도일 때, 스테로이드를 이용하여 증상을 완화하는 것이 꼭 필요하므로 점수가 대체로 0에 가까운 점수가 나타난 것으로 보인다.

5. 결 론

본 논문에서는 소셜 미디어상의 무분별한 광고를 제외한 아토피 치료법의 실제 효과를 판단하기 위해 광고를 제거하는 전처리와 효과 분석을 위한 사전 구축, 그리고 사용자 반응을 평가하는 PEM을 활용하는 소셜 미디어를 활용한 감성 분석모델을 제안하였다.

단어 빈도수와 TF-IDF 가중치를 활용해 핵심 단어를 찾아내고, 동사, 명사, 형용사 외에도 기존 감성 사전 제작에는 사용하지 않은 관형사와 부사를 사전에 포함하였다. 이를 통해 더욱 정교하게 문장 구조를 파악할 수 있었다. 이후 Word2Vec 방법을 활용하여 기존 사전에 포함된 단어와 유사한 단어를 데이터로부터 찾아내 사전 확장에 활용함으로써 사전 생성의 효율을 높일 수 있었고, 그 결과 알고리즘의 정확도가 향상되는 결과를 얻을 수 있었다. 또한, 네이버 쇼핑의 3가지 제품의 리뷰 데이터를 대상으로 기존 연구 방법과의 비교함으로써 PEM의 정확성을 확인하였고, 블로그 이외의 소셜 미디어 매체에서도 실제 반응을 파악할 수 있음에 대해 유용성을 입증하였다.

또한, '아토피 치료'와 '여드름 치료'라는 주제의 실제 데이터를 적용함으로써 화장품 제품군 이외의 대상에 대해서도 텍스트에 나타난 효과가 분석이 가능한 것을 보아, 본 논문에서 제안한 PEM은 효과를 알고자 하는 다양한 대상에 적용될 수 있을 것으로 보인다.

마지막으로 본 연구의 한계점으로 사전 구축 단계에서 네 가지 사전의 단어 개수에 편차가 존재한다는 점이 있으며 향후 연구를 통해 사전을 점차적으로 확장 및 보완해 나갈 예정이다.

References

[1] B. H. Back and I. K. Ha, "Comparison of Sentiment Analysis from Large Twitter Data Sets by Naive Bayes and Natural Language Processing Methods," *Journal of information and Communication Convergence Engineering*, Vol.17, No.4, pp.239-245, 2019.

[2] M, Ryan and S. Y. Han, "Web Scraping with Python," Hanbit Publishing Network, 2019.

[3] C. Deepti, J. Nisheeth, I. Mathur, and Y. J. Yu, "Mastering Natural Language Processing with Python," Acorn Publishing, 2017.

[4] S. M. Park, C. W. Na, M. S. Choi, D. H. Lee, and B. W. On, "KNU Korean Sentiment Lexicon - Bi-LSTM-based Method for Building a Korean Sentiment Lexicon," *Journal of Intelligence and Information Systems*, Vol.24, No.4, pp.219-240, 2018.

[5] P. S. Jang, "Study on Principal Sentiment Analysis of Social Data," *Journal of the Korea Society of Computer and Information*, Vol.19, No.12, pp.49-56, 2014.

[6] J. Y. Choi, "A Study on the Text-based Emotion Classification," Inje University Graduate School, pp.1-41, 2017.

[7] S. C. Choi, "A Study on Development of Korean Basketball League through Big Data Sentiment Analysis," Sookmyung Women's University Graduate School, pp.1-121, 2018.

[8] J. J. Lee, "Sentiment Analysis based on Korean using Recurrent Neural Network: focused on Online Movie Review," Kookmin University Graduate School Department of Data Science Data Science Major, pp.1-54, 2018.

[9] W. H. Kim, "Influencer, drug or bad?," *Korea Marketing Association*, Vol.53, No.11, pp.9-15, 2019.
[10] Y. S. Lim, S. Y. Lee, J. N. Lee, and B. K. Ryu, "An Analytical Effect Model for Atopic Therapy Using Social Media," In Proceedings of the Korea Information Processing Society, Vol.28, No.2, pp.742-745, 2019.
[11] C. W. Jun, T. Y. Choi, and J. H. Cho, "Natural Language Processing Starts with TensorFlow and Machine Learning," Wikibook, 2019.
[12] D. H. Lee, E. J. Do, J. Y. Lee, Y. Park, J. W. Oh, M. H. Lee, S. J. Hong, S. Y. Lee, J. S. Park, D. H. Nam, and H. Y. Yeom, "Multicenter questionnaires on the current management of atopic dermatitis in Korea," *Allergy Asthma & Respiratory Diseases*, Vol.4, No.4, pp.271-275, 2016.
[13] M. J. Kim, "Acne Prevention and Treatment Research," *The New Medical Journal*, Vol.42, No.3, pp.1-17, 1999.
[14] D. H. Kim, G. H. Gang, G. W. Kim, and I. Y. Yoo, "Management of Children with Atopic Dermatitis," *Allergy Asthma & Respiratory Diseases*, Vol.18, No.2, pp.39-48, 2008.

임 영 서
https://orcid.org/0000-0002-3738-8235
e-mail : dladudtj1014@naver.com
2017년 ~ 현 재 동덕여자대학교 정보통계학과 학사과정
관심분야 : Deep Learning, Big Data, Text Mining

이 소 영
https://orcid.org/0000-0002-6694-0498
e-mail : leesy970703@naver.com
2016년 ~ 현 재 동덕여자대학교 정보통계학과 학사과정
관심분야 : Machine Learning & Natural Language Processing

이 지 나
https://orcid.org/0000-0002-7925-2950
e-mail : jinalee132@gmail.com
2015년 ~ 현 재 동덕여자대학교 문헌정보학과 학사과정
관심분야 : Big data & Data Mining

류 보 경
https://orcid.org/0000-0003-0431-1546
e-mail : ryubk98@gmail.com
2017년 ~ 현 재 동덕여자대학교 정보통계학과 학사과정
관심분야 : Big data & Text Mining

김 현 희
https://orcid.org/0000-0002-7507-8342
e-mail : heekim@dongduk.ac.kr
1996년 이화여자대학교 컴퓨터학과(학사)
1998년 이화여자대학교 컴퓨터학과(석사)
2005년 이화여자대학교 컴퓨터공학과 (공학박사)
2005년 ~ 2006년 LG전자 디지털미디어연구소 선임연구원
2006년 ~ 현 재 동덕여자대학교 정보통계학과 부교수
관심분야 : Machine Learning, Deep Learning, Big Data Analysis