소셜 미디어를 활용한 아토피 치료법 효과 분석 모델

임영서*, 이소영*, 이지나**, 류보경*, 김현희*
*동덕여자대학교 정보통계학과
**동덕여자대학교 문헌정보학과
e-mail : dladudtj1014@naver.com
leesy970703@gmail.com
jinalee132@gmail.com
ryubk98@gmail.com
heekim@dongduk.ac.kr

An Analytical Effect Model for Atopic Therapy Using Social Media

YoungSeo Lim*, SoYoung Lee*, JiNa Lee**, BoKyoung Ryu*, HyonHee Kim*
*Dept. of Statistics and Information Science, Dongduk Women's University
**Dept. of Library and Information Science, Dongduk Women's University

요 약

SNS 의 발달로 이를 활용한 제품의 광고가 활발하게 이루어지고 있다. 다양한 제품군 중에서도 사용자의 피부 및 건강의 개선 효과가 나타나는 화장품, 건강보조제 등은 후기 글을 보고 실제 효과를 판단하기에 어려움이 있다. 이는 많은 양의 광고에 가려진 실질적 후기를 찾는 것이 어렵고, 포스팅의 전문을 읽는 것은 비효율적이라는 점에서 기인한다고 할 수 있다. 본 논문에서는 소셜 미디어를 바탕으로 아토피 치료법의 효과를 분석할 수 있는 효과 분석 모델을 개발하고 그 결과를 제시하였다. 먼저 많은 후기가 존재하는 키워드를 기반으로 최대 1000 개의 블로그 포스팅을 수집하였고, 광고성 글을 제외하는 자동 처리 알고리즘을 실시하였다. 다음으로 각각의 후기 글에 나타난 효과를 한눈에 알아볼 수 있도록 점수화하는 효과 분석 알고리즘을 제안하고 실험하였다. 실험결과 감마리놀렌산, 플라즈마, 락토바실러스 등이 긍정적 효과가 있는 치료법으로 나타났다. 본 논문에서 제시한 알고리즘은 제품의 효과를 점수화할 수 있으므로 아토피 치료법에 한정되지 않고, 해당 제품군인 화장품 및 건강보조제 등에 다양하게 적용될 수 있을 것으로 보인다.

1. 서론

소비자들은 상품을 구매하기 전, 더욱 많은 정보가 있는 SNS, 더 나아가 인터넷에서 사람들의 후기와 추천을 통해 제품의 구매를 결정짓고는 한다. 그러나 판매자의 광고와 해당 제품을 협찬을 받아 후기를 작성해주는 인플루언서들의 글에 가려져 실질적인 사용자들의 정보를 찾는데 어려움이 나타난다. 특히 본 논문에서 분석 대상으로 하는 화장품, 건강보조제 등은 피부 및 건강 개선 효과를 초점에 두고 광고를 하지만 실제로는 그 효과가 미비하거나 부작용이 있는 예도 있다. 따라서 해당 제품군의 효과를 판단하기 위해서는 광고성 글을 구분하여 걸러내는 것과 전반적으로 사용자들이 실제로 개선됨을 느꼈는가를 알아내는 것이 매우 중요하다. 그러나 무수히 많은 정보 중에서 광고 글을 일일이 구분하고 모두 읽는 것은 매우 비효율적이라는 문제점에 주목한다. 이에 본 논문은 치료를 위해 의약품과 더불어 화장품, 영양제 등을 사용하는 '아토피 치료법'을 주제로 새로운 알고리즘을 활용한 분석을 진행하였으며, 이를 통해 효과적인 아토피 치료법을 알아내는 것을 목적으로 한다.

본 논문에서는 뉴스와 블로그 데이터 크롤링을 통하여 6 개의 치료법을 선정해 자료를 수집하였다. 수집된 데이터를 기준으로 특정 블로그 제목과 본문 내 광고 문구를 찾아내 광고성 글을 제외하는 방식으로 광고 제거 알고리즘을 구현한다. 이후 전처리를 거친 실질적인 후기를 바탕으로 제품의 효과를 한눈에 알

아볼 수 있도록 4 가지의 사전을 활용하여 아토피 치료와 관련된 제품의 효과를 수치화하고자 한다. 기존의 감성분석에서 활용되는 감성사전은 문장의 긍정과 부정의 의미만을 구분하기에 해당 분석에는 적합하지 않다. 예를 들어 기존의 감성사전에서 '나타나다'와 같은 동사는 의미가 없지만, 화장품, 건강보조제 관련 후기 글에서는 선행하는 단어가 '증상'과 관련됐다면 부정의 의미를, '효과'와 관련됐다면 긍정의 의미를 지니는 중요한 문장이 된다. 이처럼 제품군에 따라 같은 단어임에도 새롭게 의미가 생기기도 하고, 선행하는 단어에 따라 전혀 다르게 해석되기 때문에 이에 맞는 사전을 구성하고 이를 점수화하는 알고리즘이 필요하다.

본 논문의 구성은 다음과 같다. 1 장 서론에 이어 2 장에서는 2 차에 걸친 데이터 수집과 전처리, 광고 제거 알고리즘을 소개한다. 3 장에서는 기존의 감성사전의 한계를 보완하는 4 가지의 사전을 구성하고, 이를 토대로 포스팅 별 효과를 분석하는 효과 분석 알고리즘을 제시한다. 마지막 4 장에서는 알고리즘을 활용하여 각각의 치료법의 효과를 수치화하고, 본 논문의 결론과 한계점을 제시하는 과정으로 이루어진다.

2. 데이터 수집 및 전처리

2.1 1차 데이터 수집 및 키워드 선택

크롤링에는 파이썬 BeautifulSoup 라이브러리와 request 패키지를 활용하였다. 네이버 뉴스에서 '아토피 치료'라는 키워드로 검색한 뒤, 이에 해당하는 기사 698 개의 본문 내용 전체를 수집하였다. (2019.08.07 기준)

수집된 데이터에서 2 차 수집 시 필요한 아토피 치료성분 및 치료법에 관한 키워드를 얻고자한다. 한국어 형태소 분석기인 Okt 를 활용해 명사를 추출하고, 이를 Counter 패키지와 Wordcloud 패키지를 사용하여 Wordcloud 를 생성하였다.

<그림 1> 1 차 데이터 수집 Wordcloud

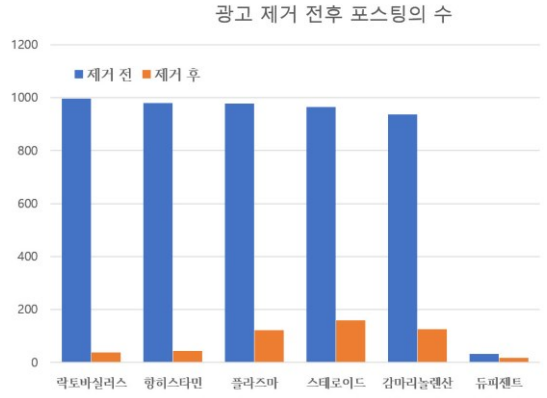
<그림 1>에서 아토피 치료성분과 치료법과 관련된 상위 20 개의 단어를 선별했다. 해당 단어들을 키워드로 하여 네이버 블로그에서 '아토피 + 키워드'의 형태로 검색하였고, 그 중 비교적 충분한 양의 데이터를 가진 '락토바실러스', '항히스타민', '플라즈마' 등의 최종 6 개의 키워드를 선택하였다.

2.2 2차 데이터 수집 및 전처리

2 차 크롤링에는 파이썬 BeautifulSoup 라이브러리와 네이버 API 를 사용하였다. 1 차 수집에서 선별한 6 개의 키워드는 '아토피 + 키워드'의 형태로 네이버 블로그 내에 검색어로 활용하였다. 각각의 키워드를 기준으로 33~1000 개의 글의 블로그 이름과 본문을 수집하였다. (2019.08.12 기준)

수집 결과 다수의 광고 포스팅이 확인되었으나 이를 일일이 제거하는 것은 비효율적이란 판단에 html 태그와 list, for 문을 활용하여 광고를 자동적으로 제거하였다. 기업의 광고를 목적으로 하는 블로그들은 비슷한 내용의 글을 반복적으로 작성한다. 이러한 경우 블로그 이름에 특정 단어가 포함되어 있다는 공통점을 발견하였다. 따라서 Okt 형태소 분석기를 활용하여 블로그명을 명사 단위로 나누고, 출현빈도가 높은 '한의원', '피부과', '코스메틱' 등의 단어를 추출하였다. 이후 블로그 이름에 해당 단어가 포함된 경우 자동적으로 포스팅을 제거하였다.

이와 다르게 개인이 블로그에서 제품을 광고하는 경우에는 반드시 '본 게시물은 **에서 제품을 무상제공 받아 작성된 글입니다'와 같은 문구를 삽입해야 한다. 따라서 블로그 내용 내에 해당 문구가 포함된 경우에도 자동적으로 포스팅을 제거하였다.



<그림 2> 광고 제외 전처리 전후 데이터 수 비교

그 결과, <그림 2> 에서 볼 수 있듯이 키워드 별 최소 46%에서 최대 97%에 달하는 광고 글이 제거되었다.

또한 광고를 제거한 데이터는 이후 사전구성시 형태소 분석이 원활하게 이루어지도록 정규표현식을 활용하여 불필요한 문자를 제거한다. 'ㅋㅋㅋ', 'ㅎㅎㅎ' 등의 단일문자 와 '^^', '~' 등의 문장기호, 이모티콘, 과도한 공백을 제거하였다.

3. 사전 구성 및 효과 분석 알고리즘

3.1 사전 구성

본 논문의 목적이 특정군의 효과를 파악하는데 있으므로 데이터의 효과적인 분석을 위하여 사전을 직접 제작하였다. 기존의 감성사전은 문장 전체의 긍정과 부정을 판단하는데 그친다. 그러나 화장품과 영양제의 효과를 판단할 때에는 긍정과 부정을 나타내는 표현 앞에 선행하는 단어가 증상이나 효과와 관련된 단어인지에 따라 그 해석이 달라진다. 뿐만 아니라 감성사전에서는 큰 의미를 가지지 않았던 표현들이 의미를 가지기도 한다. 예를 들어 '나타나다'라는 '증가'의 의미의 단어가 증상에 관련된 '건조' 등의 단어 뒤에 온다면 '건조함이 나타나다'라는 부정적인 의미로 사용된 것이다. 그러나 '수분감'이라는 효과와 관련된 단어와 함께 사용되면 '수분감이 나타나다'라는 긍정적인 의미의 문장이 된다. 이러한 특성으로 사전은 증상, 효과, 증가, 감소 총 4가지로 구성하였다.

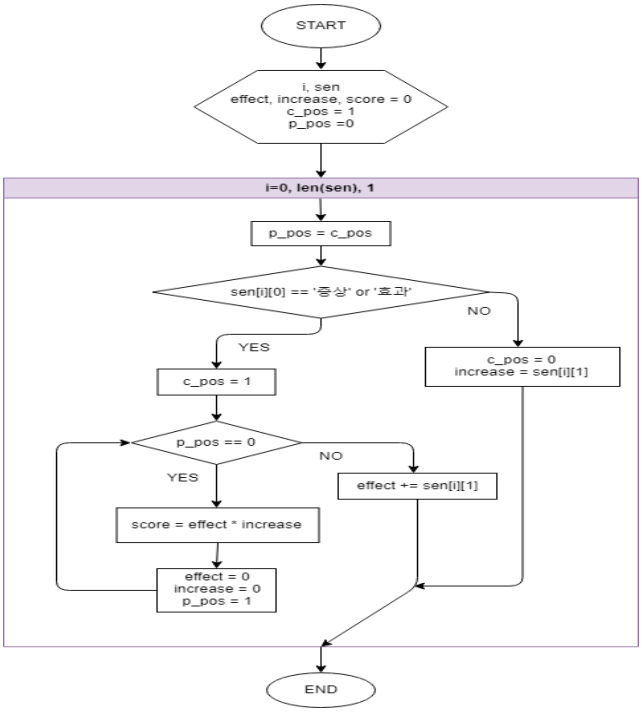
증상	효과	증가	감소
건조	촉촉	증가	감소
붉어지다	면역성	나타나다	사라지다
쓰라리다	수분감	많아지다	줄어들다
거칠다	탄력	올라오다	가라앉다
진물	효과	생겨나다	해소되다
가려움	보습	늘어나다	부족하다

<표 1> 4가지 사전 예시

<표 1>과 같은 사전의 구성을 위해 한국어 형태소 분석기인 Okt 와 Counter 패키지를 이용하였다. 블로그 데이터에서 2 글자 이상의 명사, 동사, 형용사 중 출현 빈도 3 회 이상에 해당하는 단어를 선별하고, 형태소의 변형을 고려하여 어간 형태로 추출하였다. 이후 증상, 효과 그리고 이들의 증가와 감소를 나타내는 단어를 각각의 사전으로 구성하여 텍스트 파일로 저장하였다. 증상과 효과 사전은 58 개의 단어로, 증가와 감소 사전은 50 개로 구성되어있다.

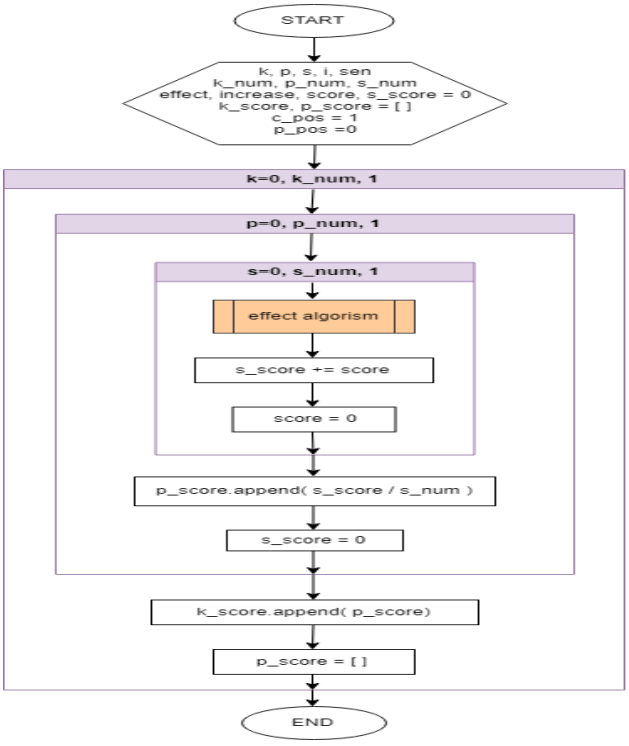
3.2 효과 분석 알고리즘

본 논문에서는 4 가지의 사전을 활용해 효율적으로 각 문장의 효과 점수를 계산하기 위하여 효과 분석 알고리즘을 제안한다. 해당 알고리즘은 하나의 문장을 기준으로 점수를 계산하는 effect algorithm 과 그 과정을 반복하여 하나의 포스팅과 키워드로 점수를 확장하는 방식으로 진행된다. 이때 각 문장은 효율적인 분석을 위하여 '가려움이 늘어났어요.'라는 문장을 [['증상,-1],['증가,1]]와 같은 리스트의 형태로 재구성한다.



<그림 3> effect algorithm

리스트로 변형한 각 문장은 <그림 3>의 알고리즘 내에서 sen 변수로 호출된다. 이를 기준으로 증상, 효과 사전에 해당하는 단어 뒤에 증가, 감소 사전에 포함된 단어가 연속될 때 각 사전의 조합에 따른 계산을 실시한다. 증상(-1), 효과(+1)를 누적하여 저장하는 effect 점수와 증가(+1), 감소(-1)를 저장하는 increase 점수를 곱한 값을 score 에 누계한다.

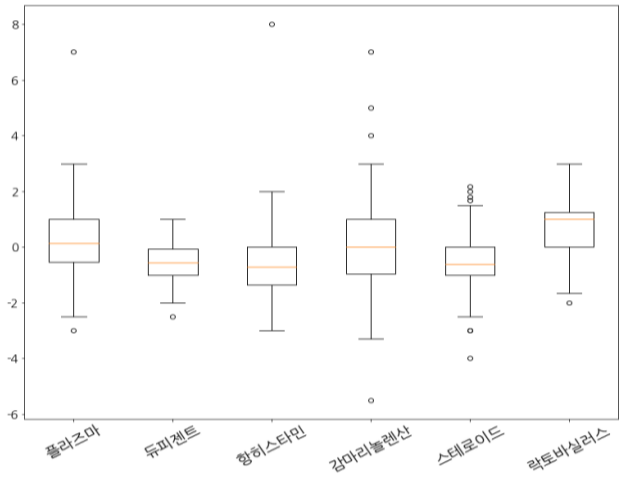


<그림 4> 효과 분석 알고리즘

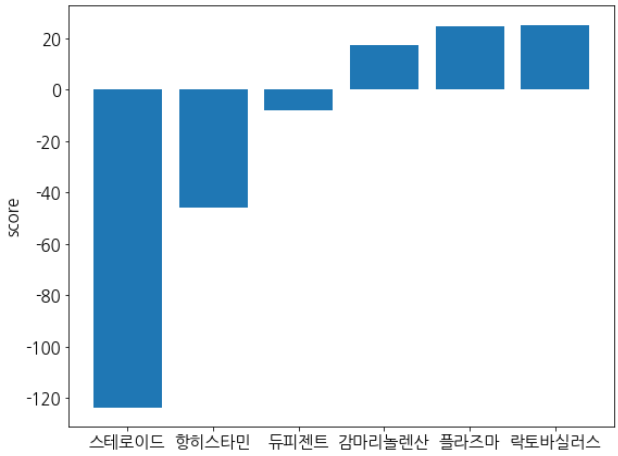
<그림 4>의 효과 분석 알고리즘은 크게 3 개의 반복문과 서브루틴인 <그림 3>의 알고리즘으로 이루어져있다. 3 개의 반복문은 각각 s(문장), p(포스팅), k(키워드) 순으로 점수의 저장이 반복된다. 각 문장의 분석이 끝날 때마다 score 는 s_score(문장 점수)에 누계된다. 그렇게 하나의 포스팅에 대한 분석이 끝나면 s_score 를 해당 포스팅에 포함된 문장의 수인 s_num 으로 나누어 변환된 값을 p_score(포스팅 점수 리스트)에 추가한다. 위와 같은 방식으로 키워드 내에 모든 포스팅의 p_score 가 계산되면 이를 k_score(키워드 점수 리스트)에 추가한다. 해당 과정들을 반복하면 최종적으로 키워드별 점수 리스트인 k_score 를 얻을 수 있게 된다.

4. 실험결과

아토피 치료법과 관련된 6 개의 키워드를 선정하여 데이터의 수집, 전처리, 사전 구성을 진행하고, 효과 분석 알고리즘을 적용하였다. 각각의 키워드의 점수 분포를 나타낸 결과는 다음과 같다.



<그림 5> 아토피 치료법의 효과 분석 분포



<그림 6> 아토피 치료법의 효과 분석 결과 분포

<그림 5>를 통해 알 수 있듯이 506 개의 포스팅 중 약 10 개의 이상치가 발견되었다.

<그림 6>은 발견된 이상치를 제거한 후, 각 키워드별 점수인 k_score 를 막대 그래프로 나타낸 결과이다. 6 개의 키워드 점수는 스테로이드, 항히스타민, 듀피젠트, 감마리놀렌산, 플라즈마, 락토바실러스의 순서로 나타났다.

5. 결론

본 논문에서는 SNS 상의 무분별한 광고를 제외한 아토피 치료법의 실제 효과를 판단하기 위해 광고를 제거하는 전처리와 사용자 반응을 평가하는 모델인 효과 분석 알고리즘을 제안하였다. 46~97%의 광고성 글을 제거한 후기들을 알고리즘을 통해 분석한 결과, 연고에 사용되는 스테로이드와 항히스타민에 대해서는 부정적인 글들이 많이 나타났으며, 달맞이꽃 오일의 성분인 감마리놀렌산, 물리적 치료기기의 물질인 플라즈마, 유산균의 성분인 락토바실러스 등은 긍정적인 후기의 비율이 더 높은 것으로 나타났다.

이를 통해 효과 분석 알고리즘이 광고를 제외한 사용자들의 실제 반응을 파악하는데 유의미한 결과를 도출해냈음을 알 수 있었다.

현재 분석 결과에서 발견되는 이상치는 향후 세밀한 형태소 분석을 통해 사전을 확대하고, 다양한 문장 구조를 파악하여 알고리즘을 보완한다면 개선될 것으로 보인다. 또한 아토피 치료법 이외에도 화장품, 건강보조제 등의 후기를 효과 분석 알고리즘을 활용해 분석한다면, 소비자들로 하여금 제품 선택의 도움을 주는 서비스를 제공할 수 있을 것으로 기대된다.

<참고 문헌>

[1] 이형호, 노기섭, “온라인 정보제공 시스템의 정보 초기 반응 상태 분석”, Vol.46, 정보과학회논문지, p.620-626, 2019

[2] 라이언 미첼, 한선용, 『파이썬으로 웹 크롤러 만들기』, 한빛미디어, 2019

[3] 딥티 초프라, 니쉬트 조쉬, 이티 마투르, 유연재, 『파이썬과 자연어 처리』, 에이콘 출판사, 2017

[4] 김유영, 송민, “영화 리뷰 감성분석을 위한 텍스트 마이닝 기반 감성 분류기 구축”, Vol.22, 지능정보연구, p.71-89, 2016