

리뷰에서의 특징 추출을 통한 특징 기반 뷰티 상품 검색 시스템

김 태 완¹ · 박 수 현¹ · 이 현 아^{2*}¹국립금오공과대학교 컴퓨터공학부 소프트웨어전공 학사과정²국립금오공과대학교 컴퓨터공학부 소프트웨어전공 교수

Feature-Based Beauty Product Search System through Feature Extraction from Product Reviews

Tae-Wan Kim¹ · Soo-Hyeon Park¹ · Hyunah Lee^{2*}¹Bachelor's Course, Department of Software, School of Computer Engineering, Kumoh National Institute of Technology, Gumi 39177, Korea²Professor, Department of Software, School of Computer Engineering, Kumoh National Institute of Technology, Gumi 39177, Korea

[요 약]

개인의 피부 유형이나 취향 등에 따라 선호 제품이 달라지는 뷰티 상품의 온라인 쇼핑에서 리뷰는 구매 의사 결정에 큰 영향을 준다. 최근 리뷰가 대량으로 축적되면서 정보의 양은 풍부해졌지만 그만큼 원하는 정보를 찾는 데 어려움이 커지고 있다. 본 논문에서는 뷰티 상품의 리뷰에서 상품의 주요 특징을 자동으로 추출하고 특징들의 극성을 시각화하는 시스템을 제안한다. 제안 시스템에서는 한국어 사전학습모델인 KoBERT를 활용하여 리뷰의 긍·부정 이진 분류와 핵심 키워드 추출을 진행하여 특징 기반 검색이 가능하다. 제안 시스템을 통해 고객은 상품이 긍정적으로 평가된 단어로 상품을 손쉽게 검색하고 상품 각각의 장단점을 한눈에 파악할 수 있다.

[Abstract]

In online shopping for beauty products, where preferences vary depending on individual skin types and tastes, reviews play a significant role in making purchase decisions. Recently, reviews have rapidly accumulated, significantly increasing the amount of information; however, this has made it difficult to find desired information. This study proposes a system that automatically extracts the key features of beauty products from reviews and visualizes the sentiment polarity of those features. The proposed system utilizes KoBERT, a Korean pre-trained language model, to perform binary sentiment classification and keyword extraction, enabling feature-based searches. The proposed system helps customers easily search for products using positively evaluated keywords and instantly identify the strengths and weaknesses of each product.

색인어 : 특징 키워드 추출, 감성 분석, 뷰티 상품, 리뷰 검색, 자연어처리**Keyword** : Feature Keyword Extract, Sentiment Analysis, Beauty Products, Review Search, Natural Language Processing<http://dx.doi.org/10.9728/dcs.2025.26.6.1721>

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Received 20 May 2025; Revised 11 June 2025

Accepted 11 June 2025

***Corresponding Author; Hyunah Lee**

Tel: +82-54-478-7546

E-mail: hallee@kumoh.ac.kr

I. 서 론

자연어는 인간의 자유로운 의사 표현의 주요한 도구로, 인터넷이 고도로 발전된 현대 사회에서는 자연어 텍스트를 이용하여 다양한 채널에서 광범위하게 소통하고 빠르게 정보를 공유한다. 특히 자연어 텍스트로 구성된 상품에 대한 리뷰는 실제 사용자의 상품에 대한 주관적인 평가와 만족도를 표현하며, 잠재적 구매자의 구매 의사 결정에 큰 영향을 끼친다. COVID-19 팬데믹 이후 온라인 쇼핑의 비중이 급격히 증가하면서[1] 리뷰의 축적이 가속화되고 리뷰의 품질 또한 향상되면서 그 영향력은 강화되었으며, 이에 따라 지엽적인 주변인의 입소문보다 폭넓은 정보를 제공하는 리뷰에 대한 소비자들의 의존도가 더욱 높아지고 있다[2].

뷰티 상품의 경우 개인의 피부 타입이나 취향, 제품 제형의 특징 등에 따라 선호하는 상품이 크게 달라질 수 있어 제품을 직접 사용한 다양한 기존 구매자의 의견이 상품 선택과 구매 여부에 큰 영향을 미친다. 이런 특성을 반영하여 뷰티 상품 쇼핑물은 리뷰 작성에 여러 혜택을 제공하여 다양한 사용자의 리뷰를 축적하고 있다. 이에 따라 제품 구매에 도움이 되는 정보가 많아졌다는 장점이 있는 반면, 확인해야 할 리뷰의 양이 방대해지면서 고객들은 원하는 정보를 찾기 위해서 리뷰를 확인하는데 더 많은 시간과 노력을 들여야 하는 어려움도 겪고 있다. 만일 사용자 리뷰에 기반하여 “순한 토너”나 “촉촉한 선크림” 같은 질의에 맞는 상품을 추천해 준다면 사용자 편의를 높일 수 있지만, 대부분의 뷰티 쇼핑물에서는 제품 이름으로만 상품을 검색할 수 있어 원하는 특징을 가진 새로운 제품을 구매해 보려는 고객의 요구에 맞지 않는다.

최근 대량의 리뷰데이터를 효과적으로 분석하고 요약하는 기술 연구가 활발히 진행되고 있다. 상품의 장단점과 주요 특징이 자연어로 표현된 리뷰를 분석하기 위해서 자연어처리(natural language processing, NLP) 연구에서는 감성 분석(sentiment analysis)을 통한 긍·부정 이진 분류(binary classification)와 키워드 추출(keyword extraction) 기법을 널리 활용한다. 자동으로 리뷰를 긍·부정 이진 분류하고 상품의 특징을 나타내는 핵심 키워드를 시각화하여 제공한다면 구매자들은 상품의 특징을 한눈에 파악할 수 있다. 이러한 연구로 리뷰를 대상으로 감성 분석을 수행한 사례[3],[4]가 있지만, 이 방식은 각 영역에 적합한 속성들을 사전에 정의해야 하는 어려움으로 데이터 수집과 레이블링에 대한 비용이 증가하고 제품 고유의 특징을 반영하지 못하는 문제점이 있다.

본 연구에서는 정제되지 않은 뷰티 상품 리뷰를 대상으로 긍·부정 이진 분류를 수행하고 분류된 리뷰에서 상품의 특징을 표현하는 핵심 키워드를 추출하여 상품을 효과적으로 검색하고 시각화하는 시스템을 제안한다. 이를 위해 자연어처리 분야에서 널리 사용되는 한국어 사전 학습된 거대언어모델(Large Language Model, LLM)인 KoBERT(Korean Bidirectional Encoder Representations from Transformers)[5]를 활

용하여 감성 분석을 수행한다. 키워드 추출 단계에서는 KoBERT 모델을 활용하여 문장과 유사도에 기반하여 문장을 대표할 수 있는 토큰을 키워드로 선정하는 방법을 제안한다. 이를 통해 기존 연구처럼 사전에 속성을 정의하지 않고 리뷰에 나타나는 사용자들이 중요하게 여기는 상품의 특징을 자동으로 추출한다. 이진 분류된 상품평과 키워드에 기반하여 구축한 상품 검색 시스템에서는 각 상품의 긍정 리뷰에서 나타나는 키워드로 상품을 검색할 수 있어 “촉촉한 선크림”과 같은 상품 특징에 대한 검색어로 원하는 제품을 찾을 수 있다. 상품 조회 화면에서는 해당 상품의 긍정 리뷰와 부정 리뷰 각각에서의 주요 키워드를 제시하여 상품의 특성을 한눈에 알아볼 수 있게 구성한다. 또한, 기존 쇼핑물에서 지원하지 않는 리뷰 검색을 제공하여 찾고 싶은 특징에 대한 사용자 리뷰만을 빠르게 확인할 수 있다.

본 논문의 구성은 다음과 같다. 2장에서는 연구에 활용된 관련 연구와 기술을 소개한다. 3장에서는 뷰티 상품에서의 특징 기반 검색의 필요성을 보이고, 4장에서는 시스템의 뷰티 상품 데이터의 수집 과정과 학습에서 사용한 레이블링된 리뷰 데이터, 화장품 성분 사전에 관해 설명한다. 5장에서는 제안하는 분석 방법을 상세히 기술하고, 6장에서는 구축된 시스템을 소개한다. 마지막으로 7장에서는 연구의 결론을 도출하고 향후 연구 방향을 논의한다.

II. 관련 연구

사용자가 작성하는 리뷰는 상품에 대한 평가가 긍정인지 부정인지로 분류할 수 있으며 이는 감성 분석 분야에서 연구되었다. 감성 분석된 리뷰에서 대표 단어들을 추출하면 기존 구매자들이 긍정적으로 평가하거나 부정적으로 평가하는 특징을 파악할 수 있으며, 이는 키워드 추출 분야에서 연구되었다. 아래에서 감성 분석과 키워드 추출 관련 연구를 살펴본다.

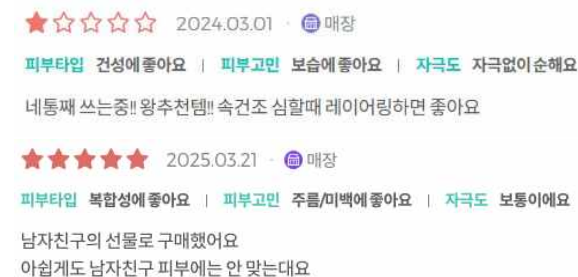
2-1 이진 분류를 통한 리뷰 감성 분석

이진 분류는 주어진 데이터를 두 클래스로 나누는 문제로 이를 해결하는 다양한 머신러닝과 딥러닝 알고리즘이 활용된다. 이진 분류의 머신러닝 알고리즘으로 가장 기본적인 접근 방식 중 하나는 로지스틱 회귀(Logistic Regression)[6]로, 입력 데이터의 선형 결합을 통해 클래스 확률을 예측한다. 이는 복잡한 비선형 관계를 학습할 때 모델 학습의 어려움이 있어, 이를 해결하기 위해 결정 트리(Decision Tree)[7]나 서포트 벡터 머신(Support Vector Machine, SVM)[8]과 같은 알고리즘이 제안되었다. 하지만 결정 트리와 서포트 벡터 머신 또한 과적합 문제와 확률 추정에 어려움이 있고 문맥이 복잡한 경우 성능이 좋지 않은 한계가 있다.

이후 등장한 Transformer 아키텍처[9]의 Encoder 구조

만을 사용한 딥러닝 모델인 BERT(Bidirectional Encoder Representations from Transformers)[10],[11]는 양방향으로 문맥을 이해하고 이를 벡터로 변환하는 것에 중점을 두어 분류 태스크에서 강점을 보인다. BERT의 장점에 기반하여 여러 다국어 모델이 연구되었으며, 이중 KoBERT[5]는 위키피디아와 뉴스 등의 대규모 한국어 데이터 세트로 사전 학습되어 다양한 한국어 자연어처리 태스크에서 기존 BERT 모델보다 뛰어난 성능을 보인다.

리뷰에 기반한 감성 분석의 기존 연구들[4],[12],[13]에서는 긍·부정 태그 대신 별점을 기준으로 극성을 판단한다. 하지만 그림 1과 같이 사용자가 수동으로 부착한 별점은 긍정 리뷰에 낮은 별점이 태그되거나 부정 리뷰에 높은 별점이 태그되는 등의 많은 노이즈가 존재한다. 추출한 상품의 특징을 나타낼 때 긍정적인 특징인지 부정적인 특징인지 판단하는 것은 중요한 요소이므로, 본 논문에서는 [14]를 기반으로 KoBERT를 활용하여 수집한 모든 리뷰에 대해서 자동 분류를 통해 레이블링한다.



*Korean is included to accurately reflect the original content

그림 1. 별점과 리뷰의 상품 평가가 일치하지 않는 예
Fig. 1. An example of a mismatch between the star rating and the review content

2-2 키워드 추출 기법

키워드 추출은 문서 내에서 중요한 단어를 식별하는 자연어처리 기술로 정보 검색, 텍스트 마이닝, 문서 요약 등에서 주로 활용된다. 기본적인 키워드 추출은 문서 내에서 중요한 단어를 직접 선택하는 방식으로, 대표적으로 TF-IDF(Term Frequency-Inverse Document Frequency)와 TextRank 알고리즘[15]이 활용된다. TF-IDF는 단어의 빈도와 역문서 빈도를 계산하여 상대적으로 중요한 단어를 선정하는 단순한 방식이다. TextRank는 그래프 기반의 알고리즘으로, 문서 내 단어 간의 연결 관계를 고려하여 주요 단어를 선정한다. 그러나 이러한 기법은 문맥을 반영하지 못하고 단어의 의미적 관계를 고려하지 않는 한계가 있다.

딥러닝 기술의 발전으로 문맥을 반영하는 키워드 추출 기법이 활발히 연구되어 최근에는 BERT 기반의 키워드 추출 기법이 널리 활용되고 있다. BERT 기반 KeyBERT[16]는 문서와 토큰을 임베딩하여 문서와 토큰의 의미적 유사성을

측정하여 키워드를 추출한다. 이는 기존 TF-IDF나 TextRank보다 문맥을 더욱 정교하게 반영할 수 있어 의미상으로 중요한 키워드를 효과적으로 선정할 수 있다. 본 연구에서는 문서의 의미적 맥락을 정확하게 반영하기 위해 KeyBERT를 활용하여 키워드를 추출한다.

III. 특징 기반 뷰티 상품 검색

기존의 대부분의 온라인 쇼핑몰은 상품명에 대한 검색을 제공하는데 상품명 기반의 검색은 새로운 상품을 찾아 사용해 보려는 사용자에게는 유용하지 않다. 본 연구에서는 상품명 대신 사용자가 기대하는 상품의 특징과 카테고리를 기준으로 상품을 검색할 수 있는 시스템을 제안한다.

뷰티 상품은 각 제품에 따른 주요 특징들이 있다. 예를 들어 스킨 케어 제품을 구매하려는 고객 중 건성 피부의 고객은 적절한 유분을 포함한 촉촉한 상품을 선호하지만, 지성 피부의 고객은 유분기 없이 수분을 채워줄 수 있는 산뜻한 상품을 선호한다. 또한 습한 여름에는 유분기가 적은 산뜻한 제품을 선호하지만, 건조한 겨울에는 보습과 유분 보충이 모두 가능한 제품을 선호한다. 하지만 그림 2와 같이 현재 서비스 중인 쇼핑몰들에서는 제품명만으로 상품을 검색할 수 있어 “산뜻한 스킨”이나 “겨울철 토너” 등의 특징을 표현하는 검색으로는 원하는 상품을 찾기 어렵다.



‘겨울철 토너’에 대한 검색결과

*Korean is included to accurately reflect the original page

그림 2. 제품 특징에 대한 검색어로 결과를 얻을 수 없는 예
Fig. 2. An example where a product feature keyword yields no search results

온라인 쇼핑몰의 각 상품에는 제조사가 제공하는 상품 상세 설명이 제공되지만, 구매자는 각 제품을 직접 사용한 사용자의 리뷰를 보다 신뢰한다. 일반 쇼핑몰의 경우 비교적 짧은 리뷰가 작성되는 것에 비해 뷰티 쇼핑몰은 리뷰 전문 사이트가 있을 정도로 리뷰의 중요성이 크고 이에 따라 상세한 사용 후기가 리뷰로 등록된다. 대표적인 뷰티 온라인 쇼핑몰 올리브영[17]의 리뷰에 대한 분석에서는 각 리뷰당 평균 27.46 어절의 리뷰가 작성되었고 100어절 이상의 리뷰는 3.42%, 80어절 이상은 8.56%로 나타나 비교적 길고 상세하게 리뷰

가 작성되는 것으로 나타났다. 이러한 뷰티 쇼핑몰의 축적된 장문의 리뷰는 많은 정보를 제공하는 장점이 있지만 각 리뷰를 일일이 확인하는데 시간과 노력이 필요한 문제점도 있다.

구매자는 희망하는 제품 특징인 ‘촉촉함’이나 ‘겨울’에 대해 긍정적으로 평가된 제품을 손쉽게 탐색하기를 원한다. 만일 특징을 표현하는 키워드로 검색했을 때 해당 특징에 대해 긍정적으로 평가한 리뷰가 많은 제품을 결과로 제시한다면 구매 편의를 높일 수 있다. 본 연구에서는 사용자가 원하는 상품의 특징(‘촉촉한’, ‘산뜻한’ 등)과 해당 상품이 속하는 카테고리(‘선크림’, ‘토너’ 등)를 기준으로 상품을 검색할 수 있는 방법을 제안하고 시스템을 구축한다. 제안 시스템은 각 상품 리뷰를 감성 분석하여 긍정과 부정으로 분류하고 리뷰에서 키워드들을 추출한다. 얻어진 긍정 키워드 중 상위 10개를 상품에 태그로 등록하여 긍정 특징으로 상품을 검색할 수 있다. 또한 긍·부정 키워드를 통해 해당 제품의 장·점을 한눈에 알아볼 수 있게 하고, 해당 키워드가 포함된 리뷰를 빠르게 탐색할 수 있도록 하여 구매 의사 결정에 도움이 되도록 한다.

4장과 5장에서는 시스템 구축에 사용된 데이터의 수집과 리뷰의 이진 분류와 핵심 키워드 추출 방식을 설명하고, 6장에서 구현된 시스템의 결과를 보인다.

IV. 데이터 수집

제안 시스템에서는 리뷰 분석과 검색 시스템 구축을 위해 국내의 가장 대표적인 헬스&뷰티(H&B) 전문 스토어인 올리브영[17]에서 상품 및 리뷰 데이터를 수집하여 사용하였다. 추가로 긍·부정으로 레이블링 되어 있는 한국경영학회의 화

장품 리뷰데이터를 분류 모델 미세조정에서 사용한다.

4-1 상품 및 리뷰 데이터 수집

데이터 수집은 스킨케어에서부터 네일까지의 올리브영의 모든 뷰티 상품을 대상으로 한다. 그림 3은 수집한 상품 정보의 일부를 보인다. 각 상품(product_id)에 대해 상품명(name), 가격(price), 할인 가격(discount_price), 구매자 리뷰(review), 리뷰 작성자 닉네임(nickname), 리뷰 작성 날짜(date), 별점(star) 등의 정보를 수집하였다. 그림 4에서는 대상 상품(product_id)별로 수집한 리뷰텍스트(review)를 보인다. 상품별 리뷰는 올리브영 웹사이트 API의 제한에 따라 최대 1,000개까지 수집하였으며, 리뷰 개수가 10개 미만인 상품은 신뢰도가 낮을 수 있어 분석 대상에서 제외하였다. 10글자 미만인 리뷰 또한 정보량이 부족하여 분석에 부적절할 영향을 미칠 가능성이 크다고 판단되어 제거하였다. 최종적으로 조건에 만족하는 5,404개의 상품과 총 272만여 개의 리뷰를 수집하였다.

4-2 레이블 된 화장품 리뷰 데이터 수집

기존의 연구에서는 별점이 붙어 있는 리뷰에서 별점을 기준으로 긍·부정을 판단하지만 그림 1에서와 같이 사용자 별점에는 노이즈가 발생한다. 이 문제를 완화하기 위해 한국경영학회에서 공개한 긍·부정 레이블을 수동 부착하고 검증한 화장품 리뷰 데이터셋[18]을 활용한다. 시스템에서는 이 데이터를 활용하여 LLM을 화장품 리뷰에 맞게 미세조정하여 4-1에서 수집된 리뷰에 대해서 감성 분석을 수행한다. 데이

id	product_id	name	price	brand	category	star	image_list	discount_price
1	A000000180474	[단독기획] 토리든 다이브인 저분자 히알루론산 토너 500ml 기획 (+화장솜 60...	29000	토리든	스킨/토너	4.8	https://image.oliveyoung....	17820
2	A000000166675	[3년연속어워즈] 아누아 어성초 77 수딩 토너 350ml 리필 기획세트(+350m...	49000	아누아	스킨/토너	4.8	http://anua.kr/dp/heartlea...	49000
3	A000000205218	[1등토너] 라운드랩 1025 독도 토너 500ml 기획 (+200ml+ 마스크팩 1매)	45000	라운드랩	스킨/토너	4.8	https://image.oliveyoung....	26400
4	A000000205155	[대용량][한정기획] 브링그린 티트리시카수딩토너 500mL+200리플(베스킨콜라...	37800	브링그린	스킨/토너	4.8	https://image.oliveyoung....	18800
5	A000000190395	[롤링진정]넵버즈인 1번 진정 맑게담은 청초토너 300ml 리필기획(+300ml 증정)	40000	넵버즈인	스킨/토너	4.8	https://image.oliveyoung....	40000
...	...	...	...	...	...	...	...	...

*Korean text is included to accurately reflect the original content of product name and brand.

그림 3. 올리브영에서 수집된 상품 데이터 중 일부

Fig. 3. Some of the product data collected from Olive Young

id	product_id	nickname	review	star	date	user_skin_info	evaluation	photo_list	recommend	useful_point
1	A000000180474	Ph5	토리든 토너색감 원정 청량미 넘침 오백미리리 다예쁘네요...	5	2024.07.19	복합성,가을형톤,다크서클,모공,민감성,주름,탄력,홍조	피부타입:건성에 좋아요,피부고민:보습에 좋아요,자극도:자극없이 순해요	2024/07/19/1721371468206.png...	0	1800.0
2	A000000180474	꼭오일밤아이태그	가성비도 좋은데 기획세트는 화장솜 같이 있어서 더 좋아요 앞도...	5	2024.07.14	None	피부타입:복합성에 좋아요,피부고민:진정에 좋아요,자극도:자극없이 순해요	2024/07/14/1720958326924.png,...	0	1512.0
3	A000000180474	Kongsuni	내 피부 - 평소엔 수분지 여름이면 지성 단독 토너만 사용시 처음엔...	5	2024.05.26	복합성,골통,다크서클,모공,민감성,잡티,홍조	피부타입:지성에 좋아요,피부고민:보습에 좋아요,자극도:자극없이 순해요	2024/05/26/1716662722662.png,...	18	1350.0
4	A000000180474	평온한하루	발써 및 통해 사용중인지 모를 정도로 만족해하며 사용하고 있습니다. 에이크...	5	2024.06.25	건성,볼륨톤,민감성,잡티,주름,탄력,트러블	피부타입:건성에 좋아요,피부고민:보습에 좋아요,자극도:자극없이 순해요	2024/06/25/1719321265876.png,...	0	900.0
5	A000000180474	yousol	물감긴한데 화장솜도 좋고 용량이 커서 든든하네요. 굿굿	5	2024.07.15	지성,혈통,모공,트러블,피지과다	피부타입:지성에 좋아요,피부고민:진정에 좋아요,자극도:자극없이 순해요	2024/07/15/1721037388777.png	0	792.0
...	...	...	...	...	...	...	...	...	...	...

*Korean text is included to accurately reflect the original content of product name, brand and review text.

그림 4. 올리브영에서 수집된 리뷰 데이터 중 일부

Fig. 4. Some of the review data collected from Olive Young

터는 그림 5와 같이 1로 레이블링 된 긍정 리뷰 7,000개와 0으로 레이블링 된 부정 리뷰 3,000개로 총 10,000개의 레이블링된 리뷰로 구성되어 있다.

	text	label
0	검은 피부엔 맞지 않네요 뽀얀 피부엔 괜찮아요 색상 조금 어두워요	0
1	봄 신상 색상 너무 저렴하게 나와서 모두 구매했는데 다 이쁨	1
2	써게 있는데 색은 그렇다 치고 잘 번져요 엄청 건조하고 썬맛 그냥 색 느낌만 만족합니다	0
3	베슬이 너무 늦고 포장 상태가 안좋아요..	0
4	페리페라 입소문 듣고 사용해본 적은 있는데 가격 저렴하게 있어서 처음 사용해보는 거...	1
...	...	...
9995	색감 너무 이뻐요 잘 쓸게요	1
9996	다 갈라지고 매트한데 색감 균일치 않습니다 실망이네 오	0
9997	커버력 괜찮고 배송도 빠르고 좋네요	1
9998	색깔 너무 이쁘고 지속력 썬	1
9999	촉촉하고 잘 스며들고 자연스럽고 너무 발림 너무 좋은 것 같아요 그리고 피부가 입술...	1

*Korean is included to accurately reflect the original content

그림 5. 레이블링된 리뷰 데이터셋 중 일부

Fig. 5. Some of the labeled review dataset

4-3 화장품 성분 사전 수집

화장품 성분은 단순한 상품의 구성 요소를 넘어서 상품 효능과 사용자 선호도에 직결되는 핵심적인 정보로, 특정 성분의 유무는 뷰티 상품의 구매 결정에서 영향을 미칠 수 있다. 화장품 성분은 일반적인 자연어처리 과정에서 틀리게 분석되는 경우가 많아, 이를 정확하게 추출하기 위해 성분명표준화 위원회에서 표준화 과정을 거친 대한화장품협회 성분 사전 [19]를 수집하였다. 시스템의 리뷰 분석과 검색에서 사용하는 형태소 분석기의 명사 사전에 수집한 성분 사전을 추가하여 성분 관련 정보를 안정적으로 인식하고자 한다.

표 1. 성분사전 표준화명칭목록

Table 1. List of standard names of ingredients dictionary

성분 코드	표준성분명		성분 코드	표준성분명
1	가공소금	...	24317	이미다졸릴메필티롤로피페라진다이온
2	가지열매추출물		24318	락티카제이바실러스
3	구멍쇠미역추출물		24319	구멍갈파래수
4	루핀아미노산		24320	괭생이모자반수
5	류신		24321	갯재미자리추출물

*Korean is used to preserve the accuracy of technical terms.

V. 리뷰 이진 분류 및 핵심 키워드 추출

본 장에서는 수집한 리뷰데이터를 긍·부정으로 이진 분류하고 핵심 키워드를 추출하는 방식을 소개한다.

5-1 리뷰 감성 분석

2-1절에서 보인 바와 같이 쇼핑몰에서 제공하는 별점에는 노이즈가 발생하며 이를 그대로 긍·부정 평가로 간주하면 적

합하지 않은 결과를 얻을 수 있어 제안 시스템에서는 리뷰 감성 분석 모델을 통해 긍·부정을 평가한다. 분류 모델은 2장에서 언급한 바와 같이 사전 학습된 KoBERT(skt/kobert-base-v1)[5],[20]를 뷰티 상품 도메인의 자연어 텍스트에 맞게 [18]의 데이터로 미세조정하여 사용한다. 학습은 이진 분류로 진행하고 카테고리는 [0(부정), 1(긍정)]로 구성한다. 하이퍼파라미터로는 batch size는 32, epoch은 4, learning rate는 0.0005, 손실함수는 cross entropy loss, 옵티마이저는 adamW를 사용한다. 입력층의 차원은 (문장길이, 768)의 선형 신경망이고 과적합을 방지하고 클래스 불균형을 완화하기 위해 drop out rate를 0.3으로 두고 bias와 layer normalization weight를 제외한 모든 파라미터에 weight decay 값을 0.01로 하여 학습을 진행하였다. KoBERT의 마지막 층에 선형 분류기를 추가하여 출력층을 구성하였다. 출력 형태는 softmax 함수를 통해 각 클래스에 대한 확률을 계산하여 더 높은 값을 가진 클래스로 결정되게 된다. 학습 데이터는 8,000개의 리뷰, 검증 데이터는 2,000개의 리뷰로 구성하였다.

이진 분류 모델의 검증과 평가에서는 해당 리뷰의 부정(0)이나 긍정(1)을 정확하게 예측한 비율을 지표로 삼아 얻어진 정밀도(Precision), 재현율(Recall), F1-Score을 표 2와 같다. [14]의 분류 모델에서 learning rate와 epoch을 조정하여 정밀도 0.9516, 재현율 0.9530, F1-Score 0.9523를 얻어 성능 향상을 보였다. 검증 데이터에 대한 혼동행렬(Confusion Matrix)은 표 3과 같다.

표 2. 성능평가지표

Table 2. Evaluation metrics

Precision	Recall	F1-Score
0.9516	0.9530	0.9523

표 3. 혼동행렬

Table 3. Confusion Matrix

Actual Class \ System Result	System Result	
	Positive	Negative
Positive	1338 (66.9%)	66 (3.3%)
Negative	68 (3.4%)	528 (26.4%)

5-2 키워드 추출

올리브영을 포함한 대부분의 뷰티 쇼핑몰에서는 카테고리별 키워드를 지정하여 그에 대한 평가를 요약하여 제시한다. 예를 들어 그림 4의 사용자 리뷰의 evaluation에서는 토너 상품에 대해서는 건성/복합성/지성에 좋은지, 피부 고민인 보습/진정/주름/미백에 좋은지, 자극의 정도가 어떠한지의 세 가지 고정된 특징에 대한 평가를 나열한다. 이처럼 지정된 키워드만 사용하면 해당 상품의 흡수력은 어떨지, 용기가 사용하기 불편하다는 등의 상품별 특이점이나 사용자가 궁금해하

는 부분들을 정확하게 파악하기 어렵다.

제안 시스템에서는 각 상품의 상품별 특징을 나타내는 키워드를 리뷰로부터 추출하여 이러한 문제를 해결한다. 키워드 추출은 리뷰 전처리와 키워드 추출로 나누어 진행한다.

1) 리뷰 데이터 전처리

리뷰에서 키워드를 효과적으로 추출하기 위해서는 키워드로 부적합한 단어들을 필터링하고 적합한 단어들의 가중치를 높일 수 있는 전처리를 진행한다. 온라인 리뷰에서는 리뷰 점수를 받기 위해 기존 리뷰를 복사-붙여넣기 하거나 동일 문구를 반복하는 경우가 있어 이러한 중복을 제거한다. 다음 단계에서는 의미 없는 정보인 특수문자나 이모티콘, 외국어, 불필요한 모음과 초성이 포함되는 노이즈 요소들을 제거한다. 이어 띄어쓰기를 교정한다. 키워드를 추출하는 과정에서는 n-gram 하이퍼파라미터를 설정하여 진행하는데, 키워드 추출 모듈로 사용하는 KeyBERT에서는 띄어쓰기를 기준으로 n-gram을 판단하기 때문에 띄어쓰기의 역할이 중요하므로 PyKoSpacing[21]을 사용하여 띄어쓰기를 교정한다.

2) 리뷰별 키워드 추출

리뷰의 전체적인 문맥을 반영하는 효과적인 키워드를 추출하기 위해 KeyBERT[16]를 활용하였다. KeyBERT는 문서에서 중요한 키워드를 추출하는데 우수한 성능을 보이는 모델로, KeyBERT의 언어 모델로는 감성 분석과 동일하게 한국어에 특화된 KoBERT를 사용하였다.

일반적인 키워드 추출에서는 명사만을 추출하는 경우가 많지만, 상품 리뷰 중 특히 뷰티 분야에서는 ‘부드럽다’나 ‘촉촉하다’와 같은 형용사가 상품의 특성을 나타내는 중요한 요소이다. 시스템에서는 화장품 상품평가에 유의미한 키워드를 효과적으로 추출하기 위해 KoNLPy[22]의 Okt(Open Korean Text)[23] 형태소 분석기를 활용하여 명사뿐만 아니라 상품의 속성을 설명하는 형용사도 키워드 후보로 선정하였다. ‘선크림’, ‘로션’과 같은 상품의 카테고리 및 상품명은 키워드 후보에서 제외하였다. 또한 뷰티 상품의 키워드는 전문 용어인 경우가 많아, 데이터 수집 단계에서 수집한 화장품 성분 사전[19]을 활용하여 Okt 명사 사전을 보완하였다.

키워드는 문장의 의미를 대표할 것으로 기대되므로 문장 임베딩과 높은 유사도를 가지는 임베딩을 가진 단어로 키워드를 선택할 수 있다. KeyBERT를 사용하여 리뷰 데이터에서 문장 임베딩과 단어 임베딩 간의 코사인 유사도를 계산하여 키워드 후보군을 추출하였다. 본 연구에서는 단일 토큰 수준에서 키워드를 추출하기 위해 KeyBERT의 n-gram 하이퍼파라미터를 (1, 1)로 설정하였으며, 이를 통해 개별 토큰의 임베딩 값과 문장의 임베딩 값을 비교하였다.

형태소 분석을 통해 얻은 단어의 어간을 키워드로 추출하여 “부드럽게”, “부드러운”, “부드럽다” 등에서 원형인 “부드럽다”를 추출하여 키워드 정규화를 수행하였다. 이를 통해 리뷰 데이터에서 상품의 평가와 특징을 나타내는 주요 키워드

들을 효과적으로 추출할 수 있었다.

표 4는 가중치가 포함된 키워드 후보군 추출의 예를 보인다. ‘순하다’와 ‘촉촉하다’와 같은 형용사와 함께 ‘뉘트’나 ‘간편하다’와 같은 신조어나 카테고리에서 사전에 정의하기 어려운 상품 각각의 특성을 반영하는 키워드가 추출된 것을 확인할 수 있다.

표 4. 가중치가 포함된 키워드 후보군 추출의 예
Table 4. Example of keyword candidates with weights

Review	Keyword
제구매만 네번 정도 했어요 일단 향이 거의 무향에 자극없고 가격도 무난해서 가성비 좋아요 세일할 때 쟁여서 쓰면 좋아요 뉘트로 무난합니다	(‘저분자’, 0.5985), (‘뉘트’, 0.5531), (‘무향’, 0.5432)
대용량이라 여름에 쓰기 좋아요 정말 자극 없이 순한 토너예요	(‘순하다’, 0.6964), (‘여름’, 0.5559), (‘자극’, 0.4955),
진짜 순하구 바르면 피부가 바로 촉촉해져요 탕글탱글 나중에 세일하면 또 살꺼예요	(‘촉촉하다’, 0.7384), (‘순하다’, 0.6712), (‘피부’, 0.6484)
순해서 가볍게 바르기 좋아요 특히 뉘트하기도 너무 간편하고 좋습니다	(‘가볍다’, 0.7407), (‘순하다’, 0.6718), (‘간편하다’, 0.5454)
양이 진짜 많아서 아직 많이 남았는데 언제 다 쓸 수 있을지 오래 쓸 거 같아서 마음에 든든하긴 합니다 자극 1도 없고 순하고 촉촉해서 좋아요	(‘든든하다’, 0.7142), (‘촉촉하다’, 0.6832), (‘순하다’, 0.6248)

*Korean is included to accurately reflect the original content

3) 상품별 키워드 해시태그

리뷰별 키워드 추출을 진행한 후, 각 상품의 모든 리뷰에서 추출된 키워드를 종합하여 해당 상품을 대표하는 주요 키워드를 선정한다. 한 상품의 여러 리뷰에서 반복적으로 등장하는 키워드를 상품의 주요 특성으로 볼 수 있으므로, 개별 리뷰에서 추출된 키워드들을 상품 단위로 집계하여 발생 빈도가 높은 상위 최대 15개의 키워드를 주요 키워드로 선택한다. 이때, 긍정 리뷰의 키워드와 부정 리뷰의 키워드를 따로 집계하고, 상품 검색에서는 긍정 키워드만을 검색에 사용한다. 그림 6은 리뷰에서 선정된 태그의 예시를 나타낸다.



*Korean is included to accurately reflect the original content

그림 6. 한 상품에서 선정된 태그의 예시

Fig. 6. Example of tags selected from a product

VI. 특징 기반 상품 검색 시스템

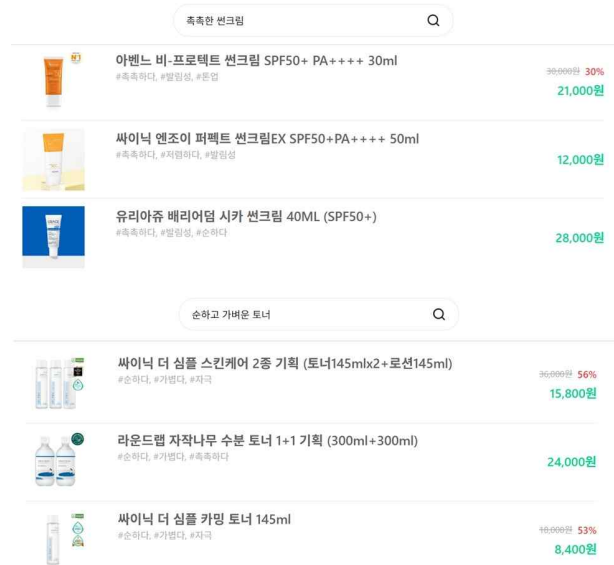
제안한 방식으로 리뷰를 감성 분석하고 키워드를 추출하고, 이에 기반한 상품 검색 시스템을 구현하였다. 시스템에서는 리뷰에서 추출한 상품별 긍정 키워드를 태그로 상품 정보에 추가하여, “촉촉한 썬크림”이나 “순하고 가벼운 토너”와 같이 상품 특징으로 검색하였을 때 사용자가 해당 상품 특징을 긍정 평가한 상품을 검색할 수 있다.

시스템 구현에서 검색 엔진은 엘라스틱서치[24]를 사용하였다. 엘라스틱서치는 역색인(reverse indexing) 기술을 사용하여 데이터의 빠르고 정확한 검색을 지원한다. 역색인은 문서 내의 각 단어를 색인하여, 검색어와 일치하는 단어를 포함한 문서를 찾아내는 방식으로 동작한다.

검색에서는 상품명 검색과 특징 기반 검색이 모두 지원된다. 뷰티 상품의 경우 상품명 및 리뷰에 미등록어가 자주 발생하여 엘라스틱서치의 Nori 형태소 분석기에도 화장품 성분 사전[22]을 추가 적용해 역색인 테이블을 구축하였다. 상품명 검색 시에는 n-gram 토큰라이저를 활용하여 입력된 검색어를 일정 길이의 연속된 부분 문자열로 분할하여 부분 일치 검색과 오타 교정, 자동완성 기능을 강화하였다. 특징 기반 검색에서는 Nori 형태소 분석기를 통해 전문 용어에 대한 인식을 높였으며, 이를 기반으로 BM25(Best Matching 25) 알고리즘[25],[26]과 네스티드 쿼리를 적용하여 정교한 검색을 구현하였다.

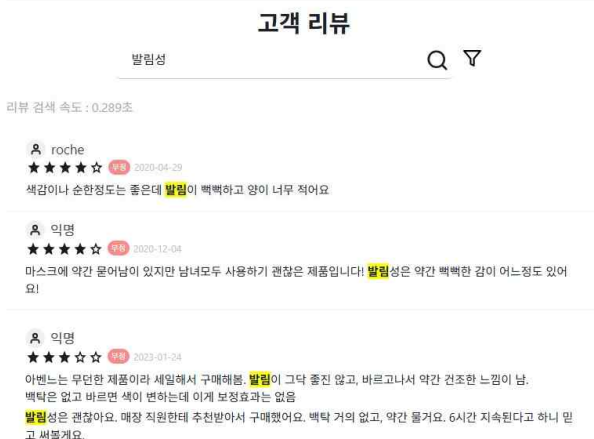
그림 7은 시스템의 검색 예시를 보인다. ‘촉촉한 썬크림’이나 ‘순하고 가벼운 토너’와 같이 “키워드(태그)+ 카테고리” 형태의 검색어로 상품을 찾을 수 있으며, 상품명 아래의 “#촉촉하다 #발림성 #톤업”이나 “#순하다 #촉촉하다 #자극 #가볍다”와 같이 각 제품에 대해 긍정 리뷰로부터 추출된 해시태그들을 확인할 수 있다.

그림 8은 썬크림, 그림 9는 토너 제품 상세 보기 페이지에서 볼 수 있는 리뷰 분석으로 얻은 긍·부정 키워드의 예를 보인다. 그림 8에서는 ‘발림성’, ‘톤업’, ‘건조하다’와 같은 썬크림 제품의 특징이, 그림 9에서는 ‘진정’, ‘촉촉하다’, ‘뉘트’와 같이 토너 제품의 특징에 맞는 키워드가 추출된 것을 확인할 수 있다. 긍정 리뷰의 상위 키워드와 부정 리뷰의 상위 키워드를 통해, 그림 8의 썬크림 제품은 ‘보정’에는 긍정적인 리뷰가 많지만 ‘건조하다’에는 부정적인 리뷰가 많음을, 그림 9의 토너 상품에 대해서는 ‘진정’의 평가는 주로 긍정적이지만 ‘불편하다’에 대한 평가는 주로 부정적임을 한눈에 확인할 수 있다. 썬크림 상품에 대해서는 ‘발림성’이나 ‘가볍다’, ‘톤업’에 대해서, 토너 상품에 대해서는 ‘자극’에 대한 평가는 긍정과 부정에 모두 포함되어, 해당 특징은 피부 특징이나 소비자 성향에 따라 각기 다르게 평가함을 알 수 있다.



*Korean is included to accurately reflect the original content

그림 7. 제안 시스템의 상품 특징 기반 검색 결과의 예
Fig. 7. Search result example based on product features in the proposed system



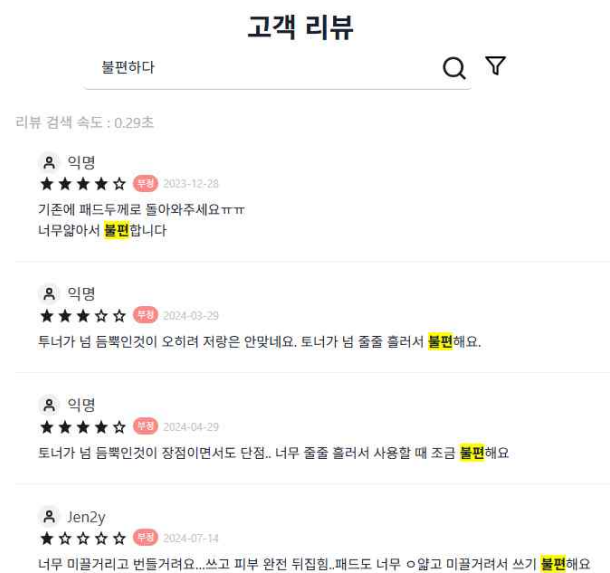
*Korean is included to accurately reflect the original content

그림 8. 썬크림 상품에서 추출된 긍·부정 키워드와 리뷰 검색 예
Fig. 8. Example of automatically extracted positive and negative keywords from reviews of a sunscreen product and review search



*Korean is included to accurately reflect the original content

그림 9. 토너 상품 리뷰에서 추출된 긍·부정 키워드와 리뷰 검색
Fig. 9. Example of automatically extracted positive and negative keywords from reviews of a skin toner product and review search



*Korean is included to accurately reflect the original content

그림 10. 상세 보기의 부정 리뷰에서의 키워드 검색 예
Fig. 10. Example of automatically extracted positive and negative keywords and review search in product details.

시스템에서는 리뷰 검색을 지원한다. 추출된 긍정적 리뷰나 부정적 부정 키워드를 클릭하면 해당 키워드를 포함한 리뷰를 바로 검색하여 제시한다. 추출된 키워드 이외의 단어도 직접 입력하여 검색할 수 있다. 그림 8의 하단에서는 ‘발림성’에 대하여, 그림 9에서는 ‘자극’에 대한 부정 키워드를 클릭한

경우의 결과를 보인다. 그림에서 별점은 4점이나 내용상 각 키워드에 대해 부정적으로 평가한 실제 리뷰 문장을 볼 수 있다. 고객은 긍·부정 키워드가 포함된 리뷰 원문을 손쉽게 확인하고 피부 특징이나 성향에 맞는지에 따라 구매 여부를 결정할 수 있다.

시스템은 상품 분류마다 사전에 정의된 고정된 특징이 아닌, 리뷰를 통해 분석한 각 상품 고유의 특징을 효과적으로 제시한다. 그림 10에서는 그림 9의 부정 리뷰의 키워드 ‘불편하다’에 대한 리뷰 검색 결과를 보인다. 대부분의 토너 상품들은 패드를 제공하지 않고 일반적인 용기를 사용하므로 모든 토너 상품에 적용되는 키워드를 사전에 정의하여 평가하는 기존 시스템에서는 각 상품에 대한 특이점을 쉽게 확인하기 어렵다. 하지만 제안 시스템에서는 각 상품에 대한 리뷰에서 키워드를 추출하므로 해당 제품의 용기나 리뉴얼된 패드의 단점을 바로 확인할 수 있다.

VII. 결 론

본 연구에서는 감성 분석된 뷰티 상품의 리뷰에서 키워드를 추출하여 상품의 장단점을 시각화하고 키워드를 상품에 태그하여 검색에 적용하는 방법을 제안하였다. 이러한 시각화와 검색으로 원하는 상품을 편리하게 찾고 대용량의 리뷰에서 추출된 상품 특징을 한눈에 파악할 수 있다.

향후 연구로는 추가적인 학습 데이터를 활용하여 리뷰 분류 시 긍·부정 뿐 아니라 중립을 카테고리에 포함해 긍정적 부정적 특징을 더 명확하게 파악할 수 있도록 모델 학습을 진행할 예정이다. 또한 뷰티 상품 도메인뿐만 아니라 전자기기, 식품 등의 다양한 온라인 판매 상품의 분야로 확장하고, 실시간으로 변화하는 쇼핑물의 상품과 리뷰들을 대상으로 해당 시스템에 적용할 수 있도록 하는 방향으로 연구를 진행할 예정이다.

감사의 글

이 연구는 국립금오공과대학교 학술연구비로 지원되었음 (2024~2026).

참고문헌

- [1] N. Shaw, B. Eschenbrenner, and D. Baier, “Online Shopping Continuance after COVID-19: A Comparison of Canada, Germany, and the United States,” *Journal of Retailing and Consumer Services*, Vol. 69, 103100, November 2022. <https://doi.org/10.1016/j.jretconser.2022.103100>

- [2] K. Bae, D. Jeong, and J. Kim, "The Effects of Product and Review Types on Review Reliance," *Journal of Consumption Culture*, Vol. 26, No. 1, pp. 39-53, 2023. <http://dx.doi.org/10.17053/jcc.2023.26.1.003>
- [3] H. Jeong, "Aspect-Based Sentiment Analysis on Cosmetics Customer Reviews," *Proceedings of the 2024 Korea Computer Information Society Summer Academic Conference*, Vol. 32, No. 1, pp. 13-16, January 2024.
- [4] S. Seo, "Sentiment Analysis of Customer Reviews Using Text Mining Techniques: Focusing on Olive Young Basic Cosmetics Review Big Data," Ph.D. Dissertation, Woosong University, 2022.
- [5] SKT KoBERT [Internet]. Available: <https://github.com/SKTBrain/KoBERT>
- [6] D. R. Cox, "The Regression Analysis of Binary Sequences," *Journal of the Royal Statistical Society: Series B (Methodological)*, Vol. 20, No. 2, pp. 215-242, July 1958. <https://doi.org/10.1111/j.2517-6161.1958.tb00292.x>
- [7] J. R. Quinlan, "Induction of Decision Trees," *Machine Learning*, Vol. 1, pp. 81-106, March 1986. <https://doi.org/10.1007/BF00116251>
- [8] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, Vol. 20, pp. 273-297, September 1995. <https://doi.org/10.1007/BF00994018>
- [9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, ... and I. Polosukhin, "Attention Is All You Need," arXiv:1706.03762, 2017. <https://doi.org/10.48550/arXiv.1706.03762>
- [10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1, pp. 4171-4186, June 2019. <https://doi.org/10.18653/v1/N19-1423>
- [11] Google-Research's Fine-Tuning Result for BERT [Internet]. Available: <https://github.com/google-research/bert/blob/master/multilingual.md>
- [12] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning Word Vectors for Sentiment Analysis," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142-150, 2011.
- [13] M.-A. Lee, Y.-J. Park, J.-Y. Na, and C.-B. Sohn, "Implementation of Review Sentiment Analysis Application Using KoBERT, KoGPT-2, and KoBART Optimized Hyperparameters," *Journal of Digital Contents Society*, Vol. 24, No. 11, pp. 2831-2840, November 2023. <https://doi.org/10.9728/dcs.2023.24.11.2831>
- [14] T. Kim, S. Park, M. Seong, J. An, and H. Lee, "Review-Based Beauty Product Feature Analysis and Search System," in *Proceedings of UCWIT 2024*, pp. 132-134, 2024.
- [15] R. Mihalcea and P. Tarau, "TextRank: Bringing Order into Text," in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 404-411, 2004.
- [16] KeyBERT [Internet]. Available: <https://maartengr.github.io/KeyBERT/>
- [17] Olive Young [Internet]. Available: <https://www.oliveyoung.co.kr/>
- [18] The Korean Academic Society of Business Administration, Online Product Review Dataset(Sentiment Analysis for Online Product Reviews), Data Repository for Business Research, 2023.
- [19] Korean Cosmetic Association Ingredient Dictionary [Internet]. Available: <https://kcia.or.kr/cid/main/>
- [20] skt/kobert-base-v1 [Internet]. Available: <https://huggingface.co/skt/kobert-base-v1>
- [21] PyKoSpacing [Internet]. Available: <https://github.com/have-n-jeon/PyKoSpacing>
- [22] KoNLPyPython Korean NLP [Internet]. Available: <https://konlpy.org/ko/latest/index.html>
- [23] open-korean-text [Internet]. Available: <https://github.com/open-korean-text/open-korean-text>
- [24] Elastic. Elasticsearch: The Heart of the Elastic Stack [Internet]. Available: <https://www.elastic.co/elasticsearch>
- [25] S. E. Robertson and S. Walker, "Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval," in *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 232-241, 1994. https://doi.org/10.1007/978-1-4471-2099-5_24
- [26] C. Manning and P. Nayak, "Introduction to Information Retrieval: Probabilistic Information Retrieval," Stanford University, 2019. <https://web.stanford.edu/class/cs276/19handouts/lecture7-probir-1per.pdf>

김태완(Tae-Wan Kim)



2020년~현 재: 국립금오공과대학교 컴퓨터공학부 소프트웨어전공 학사과정

※ 관심분야 : 자연어처리(Natural Language Processing), Large Language Model, RAG(Retrieval-Augmented Generation)

박수현(Soo-Hyeon Park)



2020년~현 재: 국립금오공과대학교 컴퓨터공학부 소프트웨어전공 학사과정

※ 관심분야 : 자연어처리(Natural Language Processing), 컴퓨터 비전(Computer Vision), 딥러닝, 패턴인식

이현아(Hyunah Lee)



1996년 : 연세대학교 컴퓨터과학과 (학사)

1998년 : KAIST 전산학과 (석사)

2004년 : KAIST 전산학과 (박사)

2004년~현 재: 국립금오공과대학교 컴퓨터공학부 소프트웨어전공 교수

※ 관심분야 : 자연어처리(Natural Language Processing), 텍스트데이터마이닝(Text Data Mining) 등