Research Article

pISSN : 2287-8920
https://doi.org/10.22716/sckt.2024.12.1.008

구매후기 토픽모델링 분석 기반 고객경험 증진 제안 연구

이한진[1], 이예은[2], 박수현[3]
[1]한동대학교 창의융합교육원 조교수, [2]한동대학교 ICT창업학부 전공 연구조교, [3]한동대학교 일반대학원 전산전자공학부 연구조교

Customer Experience Improvement Initiative based on the Topic Modeling Analysis of Purchase Reviews

Hanjin Lee[1], Yeeun Lee[2], and Suhyun Park[3]
[1]Assistant Professor, School of Creative Convergence, Handong Global University
[2]Research Assistant, Department of ICT Entrepreneurship, Handong Global University
[3]Research Assistant, Graduate School of Computer Science and Electrical Engineering, Handong Global University

[1]**Corresponding author:** discover@handong.ac.kr

Received February 16, 2024; **Revised** March 16, 2024; **Accepted** March 21, 2024

ABSTRACT

전 세계적으로 온라인 식료품 구매가 지속적으로 증가함에 따라 소비자 후기는 구매행동에 큰 영향을 주고, 이를 효과적으로 분석하면 고객경험을 향상시킬 수 있을 것이다. 이에 본 연구에서는 토픽모델링을 활용하여 고객 리뷰를 분석하고, 상품군 특성에 따른 구매동기를 파악하고자 한다. MeCab 한국어 형태소 분석 알고리즘과 불용어처리를 활용하여 국내 대표 식품구매 플랫폼에서 21~22년, 24개월간 약 120만 개의 리뷰 데이터셋을 전처리하였다. 단순한 '좋다', '만족스럽다', '추천한다'와 같은 범용어 외에 의미 있는 단어를 추출하고자 테스트 데이터셋 기반 성능을 비교하기 위해 BERTopic, LDA, CTM의 3가지 토픽모델링 유형을 적용하였다. 분석 결과, CTM은 79,616개의 유효한 데이터군에서 배송 현황, 제품 특성, 레시피 정보 등 3가지의 관련 토픽을 안정적 추출하여 가장 우수한 성능을 보였다. 이는 가치 있는 제품과 구매 통찰력을 제공하는 것을 기반으로 고객경험을 향상하는 데 학술적, 기술적 시사점을 제공한다.

As online grocery purchases continue to rise globally after pandemic, consumer reviews have gained significant influence over purchasing behavior. Analyzing user reviews in-depth and providing valuable information from the customer's perspective has become crucial for improving the customer experience. This research project aims to analyze customer reviews using topic modeling to identify product group characteristics and purchase motives. A dataset of about 1.2 million reviews from Jan. 2021 to Dec. 2022, obtained from Korea's leading food purchasing platform, was pre-processed using the MeCab Korean morpheme analysis algorithm. Researchers want to overcome the challenge of extracting such insights from general-purpose words like 'good,' 'satisfaction,' and 'recommendation.' Three types of topic modeling—BERTopic, LDA, and CTM—were applied to compare their performance on a test dataset. Finally CTM demonstrated the best performance and successfully extracted three relevant topics from the 79,616 valid review dataset: delivery status, product characteristics, and recipe-related topics. Analysis results offer academic and technical implications for improving customer experience based on providing valuable product and purchasing insights.

Keywords: Mobile grocery shopping, Combined topic modeling, BERTopic, LDA, Customer experience, Review helpfulness

© 2024 by The Society of Convergence Knowledge. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. 서 론

1.1 연구 배경

전 세계적으로 전자상거래 분야는 빠르게 성장하고 있으며 2020~2023년 팬데믹을 지나오며 많은 영역의 디지털 전환 속도가 증가하고 있다[1]. 특히, 지속적인 성장을 하며 빠르게 변화되고 있는 이커머스 시대 흐름에 따라(Fig. 1 참고), 필수 먹거리 장보기의 형태도 변화를 보인다. 2023년 오프라인 매장에서 장보는 비율이 60%에서 최근 농촌경제연구원 조사에서는 온라인 사이트 등 비대면 장보기 비율이 오히려 64%로 앞서고 있다[2]. 국내 온라인 식료품 구매율은 2010년 이후 꾸준히 증가하여 2023년 연간 약 40.7조 원(전체 약 227조원 쇼핑 거래액 중 17.9% 차지)이 거래되고 있다[3]. 신선도가 보장되는 빠른 배송과 편리한 결제, 다양한 상품 선택이 가능한 장점으로 인해 온라인 식품구매는 새로운 표준(New Normal)으로 자리 잡았다.

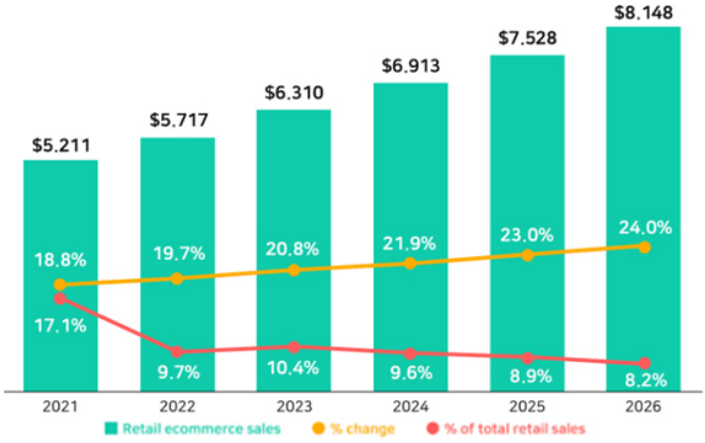


Fig. 1. 2021~2026 Global retail E-Commerce sales trend[1]

그런데도 여전히 온라인으로 신선식품 구매 시 우려되는 점은 정보의 불확실성으로 꼽힌다[4]. 신선식품의 특성상 상품의 상태와 맛, 신선도가 중요한 구매결정 요소가 될 것이다. 하지만 온라인으로 해당 상품을 구매하는 경우 이에 대해 사전에 직접 확인 및 평가가 어려운 점이 가장 큰 장애물이다. 또한, 소비자들은 신선식품을 받아본 이후에나 상품을 경험할 수 있어서 고객여정의 최종단계에서 취소, 교환, 반품, 환불을 진행하기 어렵다는 잠재적인 상황에 직면해 있다[5].

이에 신선식품과 같이 경험재적인 성격을 지닌 상품의 경우, 온라인에서 먼저 구매하여 이용한 고객들의 후기는 큰 영향을 줄 수 있다. Fig. 2와 같이 한국소비자연맹에서 조사한 결과를 보면, 실제로 온라인으로 구매하기 전 소비자들이 구매후기를 확인하

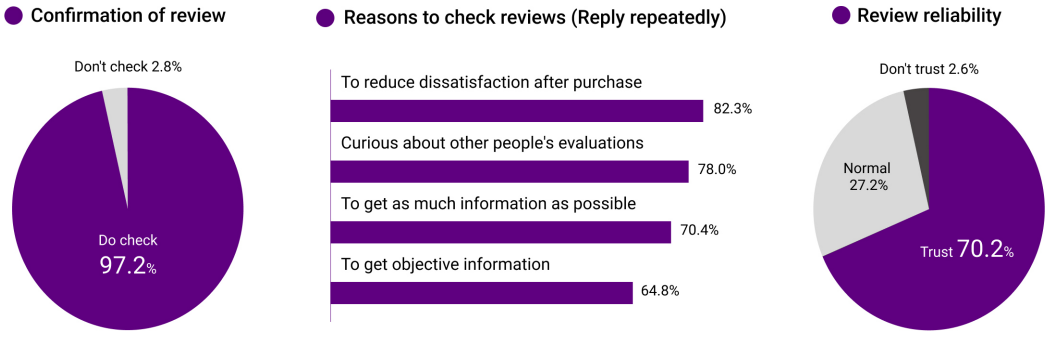


Fig. 2. Influence of customer reviews[6]

는 경우가 무려 97.2%나 된다[6]. 즉, 구매 후 불만족과 실패를 줄이기 위해, 다른 사람들이 남긴 맛과 상태에 대한 평가를 적극 참고한다고 조사되었다. 결국 온라인에서 신선식품을 구매하려는 소비자들은 충분한 상품후기 정보를 확인한 다음, 불확실성을 줄여 안심과 신뢰의 결과로 해당 상품을 구매하는 행동으로 이어진다고 볼 수 있을 것이다.

1.2 연구의 방향 및 목표

온라인 식료품 구매, 특히 신선식품의 선택에 있어 구매후기가 중요함에도 국내에서는 큰 주목을 받지 못해왔다. Open-Survey나 KPMG와 같은 산업보고서에서 단일 구매행동 중심의 소비자 인식조사를 진행해 왔거나, 농림축산식품부 산하 한국농촌경제연구원(KREI)에서 매해 식품행태조사를 자가 설문 방식으로 진행한 연구를 참고 할 수 있는 정도이다[2-3]. 아울러 실제 현업에서도 구매후기와 관련된 다양한 정보나 지표들을 크게 활용하고 있지도 못한 편으로 보인다. 국내 대표적인 온라인과 모바일 식품전문몰인 마켓컬리(Market Kurly)에서도 제공하는 구매후기에 대한 지표는 한 자리 숫자에 대부분 머물러 있는 '도움돼요'와 Best 후기(한 상품당 3개 선정)가 전부이다. 그나마 제공되는 구매후기는 검색도 되지 않으며, Best 후기도 2019년 작성일이 대부분이었다(2023.1.1 조회 기준) 이렇듯 신선식품의 대표 구매 플랫폼에서조차 해당 상품에 대한 평점, 다른 사용자들의 후기를 바탕으로 평가한 상품의 특징에 대한 정보와 같은 서비스를 제공하고 있지 않다는 점은 심각하게 보인다.

소비자들의 직접적인 피드백이 상품 구매와 해당 플랫폼 브랜드 자체 신뢰도 향상에도 효과적으로 영향을 줄 수 있다는 점은 다수의 선행연구로 지지 되었다[7-8]. 물론 탐색재와 관련된 구매후기의 연구 경향은 아마존(Amazon), 알리바바(Alibaba)와 같은 글로벌 쇼핑몰에서 구매할 수 있는 전자제품, 표준화된 생활용품 등을 토대로 이루어져 왔다[5] 또한 경험재적인 성향이 높은 여행과 숙박, 체험 서비스의 후기는 트립어드바이저(Trip Advisor), 에어비앤비(AirBnB) 사이트의 데이터를 토대로 진행된 연구가 다수 소개되었다[9-10]. 이에 반해 데이터 기반의 식품구매후기에 대한 심층적인 분석은 상대적으로 부족하여 학술적, 실무적, 정책적인 간극을 줄이고자 한다.

이에 본 연구에서는 불확실성을 줄이고 고객경험을 향상할 수 있는 온라인 식품전문몰의 구매후기를 토픽 모델링 방법론을 통해 분석해 보고자 한다. 즉, 국내 대표 식품거래 플랫폼의 후기 데이터를 수집하여 전처리 후 적합한 모델링을 통해 주요한 정보를 고객 관점에서 제공해 보고자 한다. 나아가 이러한 분석 데이터가 소비자에게 어떤 가치를 전할 수 있을지 실제 앱 서비스 화면을 구현하여 시연하는 과정까지 진행하고자 한다. 본 연구 수행을 통해 소비자들은 더 합리적인 소비를 할 수 있는 효용을 가질 수 있으며 플랫폼은 고객경험을 향상해 좋은 서비스-브랜드-제품-소비자의 유대관계를 구축할 수 있을 것이다.

2. 선행연구 검토

2.1 이론적 배경

보편적으로 구매후기는 특정 상품을 구매하거나 서비스를 이용한 소비자들이 자신의 사용 경험이나 평가를 댓글의 형태로 온라인 게시 공간에 게재한 글이다[11]. 오늘날 90%의 소비자들이 구매 결정 시에 후기를 읽는다는 조사 결과가 있는 만큼 구매후기는 중요하고 많은 영향을 주고 있다[8,12]. 다수의 선행연구는 기본적으로 후기 속성에 대해 평점과 총 개수로 구분하여 기업의 성과 창출에 미치는 영향에 관해 연구하였다[13-15]. 보통 구매후기는 주로 서술적인 형태로 나타나지만, 평점과 같이 점수화로 표현하거나 이미지, 영상 등이 함께 제시되는 경우도 많다[16]. 이에 구매후기 속성(평점, 댓글의 양, 이미지 정보)은 디지털 시대의 대표적인 소비자 참여 형태라 할 수 있다[17].

소비자행동이론(Consumer Behavior Theory) 관점에서 구매후기는 구매행동(Funnel)의 상위 단계 중 '인지(Awareness)'와 '고려(Consideration)' 측면에서 중요한 역할을 하는 것으로 연구되었다[2]. 구매후기를 구성하는 주요한 속성인 평점이나 댓글

의 규모는 기존 구매 고객들의 구체적인 구매행동의 결과로 인식되어 긍정적인 태도를 형성하고[7], 적극적인 고객여정을 연결한다[9]. 나아가 이미지 정보가 함께 포함된 후기의 풍부한 내용과 후기의 유용성은 구매단계(Funnel)에서 높은 전환(Conversion)을 끌어낸다는 다수의 연구결과도 어렵지 않게 찾아볼 수 있다[14,19]. 이는 온라인 거래를 둘러싼 생태계 속의 고객뿐만 아니라, 제품의 생산자, 상품 판매자, 서비스 기획자, 마케터, 서비스 개선 엔지니어, 고객센터 담당 등 여러 이해관계자에게 중요한 구매여정의 경험적 토대를 제공한다고 밝혀졌다[5].

한편 구매후기 유용성은 온라인 상품 후기가 다양한 상품에 대한 실질적인 정보(품질, 사용 경험 등)를 제공하여 다른 소비자들이 구매 결정을 내릴 때 도움을 주는 능력을 의미한다는 점도 확인되었다[9-10]. 특히, 온라인 B2C 쇼핑 사이트는 일반적으로 소비자가 후기의 유용성을 평가할 수 있는 투표 도구를 제공한다. 후기를 읽은 후 독자는 '좋아요(Like it)', '도움 돼요(Helpfulness)' 등의 버튼을 클릭하여 후기 유용성에 투표할 수 있다. 후기 유용성은 제품 품질의 진단성을 반영하며 "제품을 평가하는 데 도움이 됨", "제품에 익숙해지는 데 도움이 됨" 또는 "제품을 이해하는 데 도움이 됨"으로 측정되었다[17]. 따라서 후기 유용성은 종종 온라인 상품 구매 의사결정을 지원하는 능력의 주요 척도로 채택된다[18-19].

이렇듯 급변하는 산업과 세분화된 디지털 서비스가 출시되며 구매후기 연구의 지평이 넓어지고 있다. 그런데도 대부분의 데이터 분석 기반의 연구는 구매후기의 규모와 효과에 초점을 맞추고 있어, 분야별 구매후기의 다양한 측면에 관한 심층적인 연구는 제한적이었다[7]. 이에 본 연구에서는 식품구매에 있어 구매후기를 체계적으로 수집하여 선행연구에서 검증된 분석 틀을 활용, 고객과 생산자, 판매자와 플랫폼의 주요 이해관계자들에게 의미 있는 분석결과를 제공하고자 한다.

2.2 분석 방법 설계

전체적인 데이터 분석은 선행연구를 종합적으로 검토하여 다음의 5단계로 진행할 것이다[20-21]. 먼저 데이터 수집 및 데이터 구조의 기초적인 탐색 후 분석이 가능하도록 전처리를 수행할 필요가 있다. 이후 적합한 알고리즘을 사용하여 토픽 모델링을 구체적으로 수행하고[22], 유효한 토픽 키워드를 추출하는 단계를 거친다. 마지막으로 추출 결과를 적절한 수의 키워드로 시각화하고[23], 이를 기반으로 프로젝트에서 추구하고자 했던 데이터 분석의 내용과 연관지어 해석하고 평가하는 과정을 수행한다.

1) 데이터 수집 및 구조 : 우선 연구 대상으로 선정한 플랫폼의 구조를 파악하고, 이곳에서 적합한 방법으로 데이터를 수집한다. 파이썬(Python)의 Selenium, Beautiful_soup 라이브러리를 활용하여 구매후기 데이터군(Dataset)을 수집하고, 이 데이터셋 안에 구매후기와 관련된 다양한 요인(Feature)들이 어떤 형태로 포함하고 있는지 확인할 것이다[20].

2) 데이터 전처리(Pre-process): '후기 내용'의 주요한 피처를 전처리하여 토픽 모델링에 적합한 형태로 가공하는 단계를 거친다. 데이터 분석 관련 선행연구에 근거하여 유효한 데이터를 추출하기 위해 불용어사전 기반 오타, 특수문자, 이모티콘 및 반복적 기호 제거, 특정 표현이 반복되는 문장을 제거하는 등의 방법을 적용할 수 있다[23].

3) 토픽 모델링(Topic Modeling): 전처리된 후기 내용을 기반으로 토픽 모델링을 수행할 것이다. 후처리 토큰(Token)화 및 품사 태깅(Tagging)기준 검토하는 과정을 통해 후기들이 어떤 토픽으로 구성되어 있는지 추론한다. 본 연구에서 검토할 토픽 모델링은 LDA, CTM, BERTopic 알고리즘이며, 각 모델링의 결과에 따라 더 좋은 결과의 모델을 사용할 것이다[24].

4) 토픽 키워드(Topic Keyword) 추출: 토픽 모델링 결과를 기반으로 각 토픽의 키워드를 추출한다. 키워드는 각 토픽이 대표하는 주제를 잘 나타내는 단어로서, 해당 상품에 대한 구매후기 내용을 한눈에 볼 수 있도록 도움을 준다[5].

5) 키워드 시각화 및 시사점 도출: 마지막으로 주요 키워드의 시각화를 통해 소비자행동 관점에서 구매여정에 있는 고객에게 가치 있는 정보를 제시하려고 한다[2,5]. 나아가 구매고객 관점의 시연 서비스 활용을 통해 사용자로부터 평가와 검증을 받을 수 있는 체계의 시사점과 함의를 검토할 예정이다.

3. 연구 방법

3.1 탐색적 데이터 분석(EDA)

2023.01.01 기준 국내 대표 식품전문몰(마켓컬리 MarketKurly)의 베스트상품 90개에 대한 상품별 댓글 데이터를 파이썬(Python) 웹 크롤링 라이브러리 셀레니움(Selenium)을 활용하여 약 1,955,014건을 수집하였다. Table 1과 같이 수집된 csv 형태의 데이터는 상품코드, 후기 베스트 여부, 후기 내용, 후기 글자 수, 사진 포함 여부, 사진 개수, '도움돼요' 개수, 후기 등록 일자, 후기 등록 경과일, 상품명, 상품명 글자 수, 브랜드명, 카테고리 등 총 13개의 항목으로 확인하였다.

Table 1. Food purchase review data frame (Sample)

후기 번호	구매자 등급	베스트 여부	후기내용(Review)	후기 글자수	사진 개수	도움 돼요	등록 일자	할인율	소구문구	상품 특징	배송 형태	카테 고리	구매 혜택
1	프렌즈	일반	감자 자주 구매하는데, 요번엔 알 사이즈가 각기 다르네요	39	1	0	2022-12-16	10%	[KF365]믿고 먹을 수 있는 상품을...	Kurly Only	샛별 배송	채소	무료 배송
2	퍼플	베스트	가격이 좀 있기는 하지만, 싱싱하고 다양하여 요리에 사용...	127	3	2	2022-03-23	10%	[KF365]믿고 먹을 수 있는 상품을...	Kurly Only	샛별 배송	채소	무료 배송
...	...	...	...	...	...	...	...	...	...	...	...	...	...

수집된 전체 데이터세트의 메타(meta)정보에 해당하는 속성값 13종을 Type 별로 정리하면 크게 4개 유형으로 나누어 볼 수 있다. 먼저 string 타입으로 수집된 후기 데이터는 제품 카테고리, 브랜드명, 후기내용 등 5종의 속성값이 분류되었고, 후기 글자 수, 사진 개수 등 int/float는 5종이 뒤를 이었다. 해당 여부(Y/N)로 분류되는 bool은 후기 베스트 및 사진 포함 여부의 속성값이, 후기 등록일자는 date 데이터 속성값으로 각각 분류할 수 있었다. 아울러 결측치는 없는 것으로 확인되었다.

수집된 후기가 작성된 기간은 2019.10부터 2022.12까지였으나 선행연구를 검토한 결과, 신선식품의 다양한 특성상 생산 시기, 판매 가능 기간, 원산지의 작황 상태, 농산물 가격변동에 따른 판매 상태, 재고와 물량, 신선도를 고려하여 최신 후기를 선정하는 것이 필요하다고 판단하였다[2,5]. 이에 후기의 최신성을 확보하면서도 충분한 후기 게시글 노출을 고려할 때, 2021년 1월부터 2022년 12월까지 총 24개월의 1,026,780건 데이터 세트를 분석에 활용하고자 선별하였다.

아래 Fig. 3은 전체 데이터를 이용하여 가장 많이 나타난 상위 20개의 빈도수를 나타낸 그래프이다. 상위 20개 단어가 전체 토

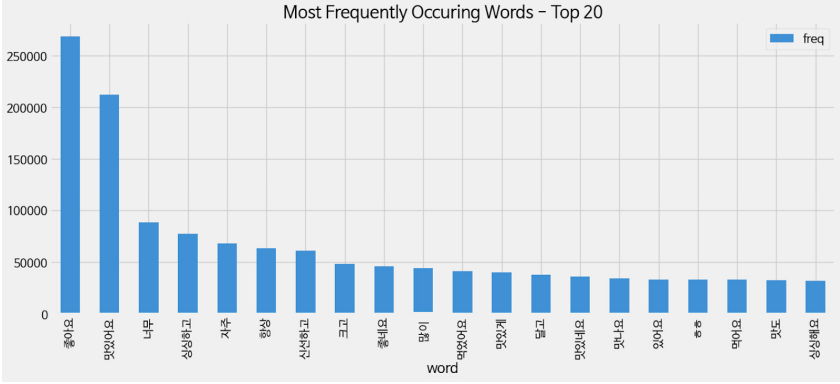


Fig. 3. Top 20 words by posting frequency in the entire dataset

큰 중에서 차지하는 비율은 약 59.3%였는데, '좋아요', '맛있어요', '싱싱하고', '신선해요'와 같이 대체로 긍정적인 단어가 많이 사용되었음을 알 수 있다. 이는 긍정 경험을 후기로 남기는 충성 고객행동의 일환으로 해석할 수 있었다.

3.2 대표 분석대상 설정

그러나 90개의 상품별로 고빈도 출현단어의 분포가 유사하여 상품별 구매후기의 특징이나 상품군의 차이를 살펴보기 어려웠다. 기존 후기 연구에서도 전체 후기를 대상으로 분석하여 '좋아요' '맛있어요' '추천'과 같은 범용적인 단어를 추출한 해석의 한계와 같이 실제로 상품군의 특징과 구매동기를 파악하기가 어려웠다[7-8]. 본 프로젝트에서는 고객후기 데이터를 토픽모델링 기반으로 분석하여 연관도가 높은 키워드를 중심으로 소비자들의 구매여정에 접목 후 실질적인 가치를 제공하고자 했다. 이에 온라인 식품 쇼핑몰 전문몰의 채소 카테고리 내에서 가장 많은 후기(전체 중 14.9%)를 보유한 대표상품 '감자(상품명: [KF365] 감자 1kg)'를 분석대상으로 선정했다.

세계식량농업 통계연감에 따르면, 감자는 전 세계적으로 생산량과 소비량이 증가하고 있으며, 국내에서도 1인당 연간 감자 평균 소비량이 약 50kg에 달하는 것으로 조사되었다[24]. 그 위상에 걸맞게 신선식품의 대분류 즉, 양곡, 채소 및 과일(농산), 고기와 계란(축산), 해산물(수산) 등의 큰 카테고리 중 베스트 상품의 가장 높은 추천 순위를 차지하고 있다. 왜냐하면 일반적으로 감자는 채소 분류 중 연중 상시 구매할 수 있어 보편적으로 섭취되고 쓰임새가 많은 특징을 가지고 있기 때문이다. 이에 해당 식품 전문몰 구매 플랫폼의 초창기부터 대표적인 제품으로 자리잡아 거래됐으며, 따라서 상당한 규모의 판매량에 따른 구매후기와 다양한 후기내용을 포함하고 있다. 대표적인 저장성 식재료이면서 활발한 거래량을 보이며, 다양한 요리 방법으로 섭취할 수 있다는 장점으로 고객들의 사랑을 받는 식품의 구매후기 분석은 충분한 가치가 있을 것이다.

이렇게 전체 후기 데이터 중 감자와 관련한 후기를 추출한 결과, 총 97,000개의 분석 가능한 데이터 세트가 1차적으로 확인되었다. 아래 Fig. 4에서는 등록된 구매후기 글자 수의 규모를 나타내는데, 최소 3자부터 최대 613자까지의 분포를 보였으며 후기당 평균 길이는 17.2자였으며 표준편차는 약 14.7이었다. 아울러 주요 분석대상 구매후기에 가장 많이 사용된 단어 상위 20개 중 위에서 살펴본 전체 구매후기 데이터의 상위 단어와 18개가 중첩되었다. 고유명사인 대표상품명 외 전체적으로 '좋아요', '맛있어요', '싱싱하고'와 같은 표현이 동일하게 다수 사용되었음을 볼 수 있었다. 대표상품의 전체 토큰 중 상위 20개 단어가 차지하는 비율은 약 57.8%로 나타났다.

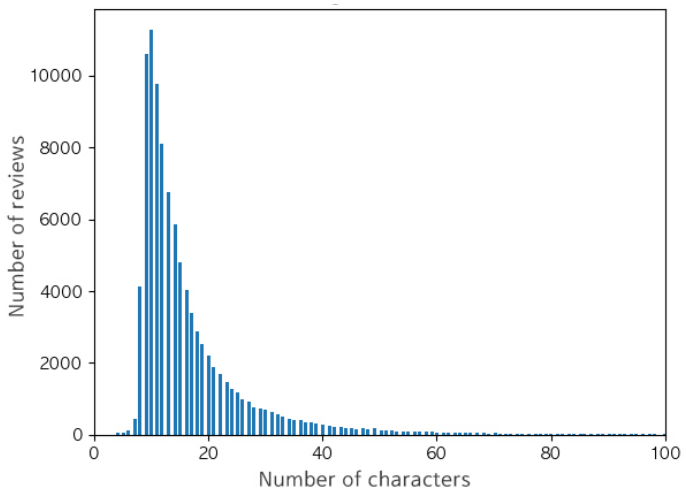


Fig. 4. Distribution of purchase review length

3.3 데이터 전처리

본격적인 분석을 위해 유효한 데이터를 추출하기 위한 데이터 전처리 과정을 거쳤다. 전처리 과정은 크게 단어와 문장 차원에서 검토한 접근이 이루어졌는데, 단어 차원에서는 특수문자, 이모티콘(Emoji) 및 반복적 기호 제거, 문장 차원에서는 영어 텍스트 및 특정 표현이 반복되는 문장을 제거하였다[21]. 특히, 유효한 분석이 어려운 특수문자, 이모티콘, ‘ㅋㅋㅋ’, ‘ㅠㅠ’와 같은 반복적인 기호는 데이터에서 제거하였다. 이러한 요소들은 긍정 혹은 부정적인 정서를 가지고 있지만, 엄밀한 기준으로 표준화하는 것이 어렵기 때문에 이번 분석에서는 우선 제외했다. 이에 단어 차원의 전처리로 문장 기준의 데이터 개수에는 차이가 없었으나 단일 문장 데이터별 평균 글자 수가 17.54에서 17.51로 0.3 글자 줄어들었다.

다음으로 데이터 탐색 단계 과정에서 발견된 후기 데이터 중 문장 대부분이 영어로 작성된 부분도 삭제하는 과정을 거쳤다. 해당 문장에는 식품 플랫폼 자체를 의미하는 ‘Kurly’와 URL, 무의미한 영문 오타가 포함되어 본 연구의 분석 목적에 적합하지 않다고 판단하였다. 이 문장 차원의 전처리로 11개의 데이터가 제거되어 분석가능 데이터 세트는 총 96,989개로 필터링되었다. ‘조아조아조아조아’, ‘마시따마시따마시따!’ 등 비표준어 반복 표현이 있는 문장들도 분석 결과에 좋지 않은 영향을 줄 수 있는 노이즈로 간주하여 삭제 처리했다. 문장에서 단일 토큰이 전체 토큰의 30% 이상을 차지하는 경우 의미 없는 반복으로 간주하여 제거할 수 있다는 선행연구를 따랐다[21]. 이러한 반복 표현이 포함된 문장 17,373개의 데이터가 제거되어 최종적으로 79,616건(82.1% 잔존)의 필터링된 유효한 후기 데이터 세트를 기반으로 토픽 모델링을 진행하였다.

4. 연구 결과

4.1 모델링 적용 결과

4.1.1 BERTopic

우선 가장 널리 활용되는 BERT(Bidirectional Encoder Representations from Transformers)는 구글의 Devlin(2018)이 제안한 딥러닝 모델로, 상대적으로 적은 자원으로도 아주 빠르고 좋은 성능을 내는 자연어 처리기법(NLP)으로 활용되고 있다[22]. 즉, 사전 학습된 대용량의 레이블링 되지 않는 데이터를 이용하여 언어 모델을 학습하고, 이를 토대로 특정 작업(문서 분류, 질의응답, 번역 등)을 위한 신경망을 추가하는 전이 학습법의 일종이다. BERTopic은 BERT 기반 토픽 모델링 알고리즘으로, 텍스트

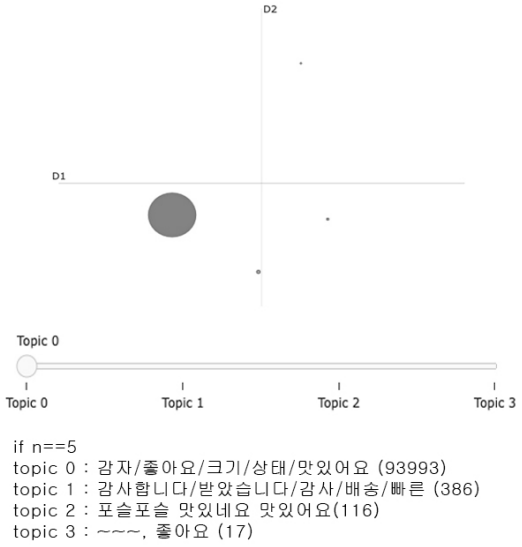
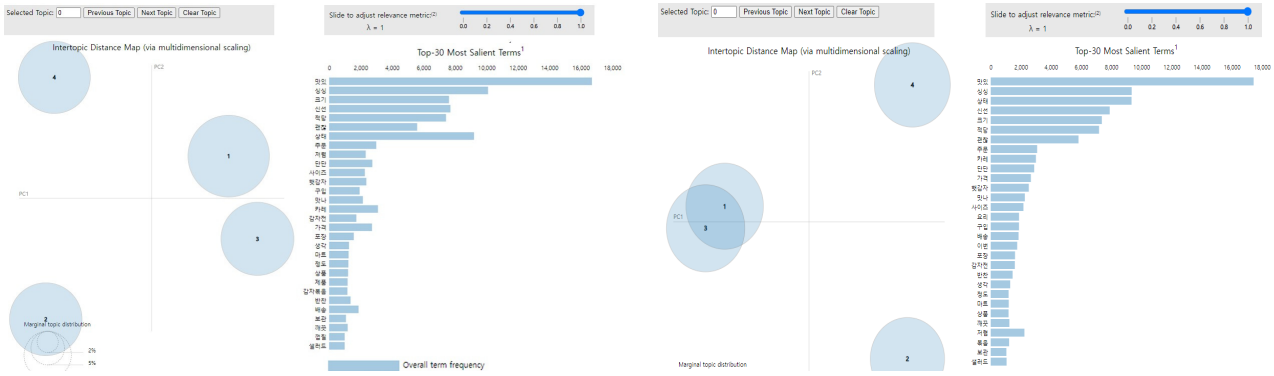


Fig. 5. BERTopic modeling result sample

데이터에서 의미론적으로 관련된 문서들을 클러스터링하여 토픽을 추출하는 방법이다[20]. 그러나 Fig. 5과 같이 여러 개의 토픽을 선택했을 때 한 토픽에 단어들이 크게 몰리는 경향이 있었으므로 이번 모델 후보에서 제외했다.

4.1.2 LDA(Latent Dirichlet Allocation)

LDA(잠재적 디리클레 할당) 알고리즘은 텍스트 데이터에서 토픽을 추론하는 통계적 모델로, 단어의 출현 빈도와 토픽의 분포를 결합하여 각 문서의 토픽 분포를 추정하는 방식이다[23]. 토픽 수를 입력받으며 전체 문서에 토픽을 랜덤으로 할당한 후, 토픽의 재할당을 반복 수행하여 문서와 단어의 적합한 토픽을 찾는다. 이를 통해 유사한 토픽을 가진 문서를 그룹화하고, 텍스트 데이터의 구조를 쉽게 파악할 수 있다. 그러나 무작위로 토픽이 할당되므로 분석 결과의 재현성이 떨어지기 때문에 같은 데이터로 같은 분석을 진행했더라도, 결괏값이 매번 달라질 수 있다. 아래 Fig. 6처럼 같은 데이터 세트에서도 매번 완전히 결과가 다르게 나오는 점을 확인한 결과, 분석 과정의 재현성이 떨어져 최종 모델에서 제외했다[23].



(a) (b)

Fig. 6. (a) LDA #1 Analysis result and (b) LDA #2 Analysis result sample

4.1.3 CTM(Combined Topic Modeling)

CTM(조합된 토픽 모델링)은 단어의 분포 패턴과 문서 간의 유사도를 함께 고려하여 토픽(분류 주제)을 추론한다[21]. 전통적인 토픽 모델링 방법(BERTopic, Latent Dirichlet Allocation 등)은 단어의 출현 빈도에만 기반하여 토픽을 추론하지만, CTM은 단어의 출현 빈도 외에도 단어의 공간적 위치 정보를 고려하여 토픽을 모델링한다[22]. Fig. 7에서 제시된 바처럼 CTM의 주요 알고리즘은 크게 2개의 수식으로 구현되는데, 먼저 현저하게 드러나는 검출감지성 혹은 돌출성(Saliency)은 생성되는 토픽을 결정하는 데 특정 용어 w가 얼마나 유익한지를 나타낸다[25]. 또한 관련성(Relevance)은 사용자가 토픽을 해석하는 데 유용한 순서대로 용어의 순위를 유연하게 지정할 수 있도록 설계되어있다[25]. 이렇게 구현되는 CTM은 재현성과 토픽 분포에서 타 모델 대비하여 우수한 결과를 나타내기에, 본 데이터 분석 프로젝트의 최종 적용모델로 선정했다.

$$saliency(term\ w) = frequency(w) * [sum_t p(t|w) * \log\left(\frac{p(t|w)}{p(t)}\right)]$$

$$relevance(term\ w \mid topic\ t) = \lambda * p(w|t) + (1-\lambda) * \frac{p(w|t)}{p(w)}$$

Fig. 7. Saliency and relevance formula of CTM[25-26]

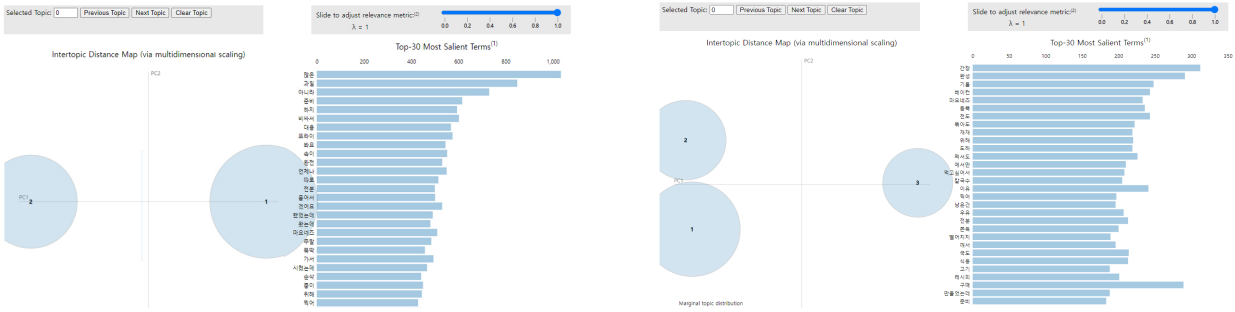
4.2 선택학습모델 후처리

CTM은 문장으로만 데이터를 입력받을 수 있기 때문에 텍스트 데이터를 모델링한 후, 나온 결과를 후처리하여 유의미한 단어들을 추출하는 과정이 수행된다[21]. 후처리 단계는 CTM의 효과와 해석의 가능성을 향상하기에 다음의 단계를 거쳐 진행할 것이다. 먼저 토큰화(Tokenization)를 진행하는데, 이는 주어진 텍스트 데이터를 단어나 구문으로 나누는 과정을 말한다[21]. 이러한 분석을 위해 한국어 형태소 분석기인 MeCab(메캅) 라이브러리를 사용하여 형태소를 분석하고, 품사 태그를 추출했다. 이 과정을 품사 태깅(Part-of-Speech Tagging) 과정으로 부르며 이를 통해 토픽 분석에 기여하는 실질적 의미가 있는 단어에 집중했다. 나아가 유의미한 단어 추출을 위해 일반적으로 어근을 중심으로 한 명사와 부사 품사를 선택하는 경우가 많은데, 본 연구에서는 탐색적 데이터 분석(EDA)을 기반으로 다음 7종의 품사 태그들만 선택하기로 했다[20]. 즉, ‘XR’ (어근), ‘NNG’ (일반명사), ‘NNP’ (고유명사), ‘NNB’ (의존명사), ‘MAG’ (일반부사), ‘MAJ’ (접속부사), ‘MM’ (관형사)를 선택했다. 단, 고유명사 중 ‘컬리’와 ‘KF365’(자체 상품명 일부) 등의 단어는 제외함으로써 모델링 과정에서 비실질적인 용어들을 걸러냈다.

다음 단계로 토픽별 단어 개수를 분석하기 위해 널리 쓰이는 파이썬 컬렉션 모듈(Python Collection Module)의 ‘카운터 클래스(Counter Class: CC)’를 활용했다[21]. 본 클래스는 Key와 Value를 한 쌍으로 저장하는 자료형인 사전 형태로 반환하여 해쉬(Hash) 가능한 객체들을 계산하는 딕셔너리(Dictionary)이다. 이에 카운터 클래스(CC)를 이용하여 토픽별 단어들의 개수를 계산하고 상위 15개의 가장 높은 빈도 단어들을 추출하였다. 그 결과 ‘적당’, ‘싱싱’, ‘신선’, ‘가격’, ‘볶음’, ‘카레’, ‘크기’, ‘양도(많고)’, ‘너무’, ‘포슬포슬’, ‘자주’, ‘이번’, ‘알도’, ‘단단’, ‘습기’라는 목록을 확인할 수 있었다. 특히, ‘적당’, ‘신선’ 등 각 토픽에 중복으로 들어가는 단어들의 경우, 해당 단어의 빈도가 더 큰 쪽의 토픽으로 할당하는 조정 과정을 거쳤다. 이 과정을 통해 각 토픽과 관련된 주요 특징과 경향에 대한 명확한 통찰을 얻을 수 있었다. 또한 각 토픽을 일반명사, 고유명사, 형용사, 어근, 일반부사, 접속부사, 관형사, 의존명사로 추출한 상위 15개 단어의 비율은 각각 27.2%, 21.3%, 19.5%로 집계되었다.

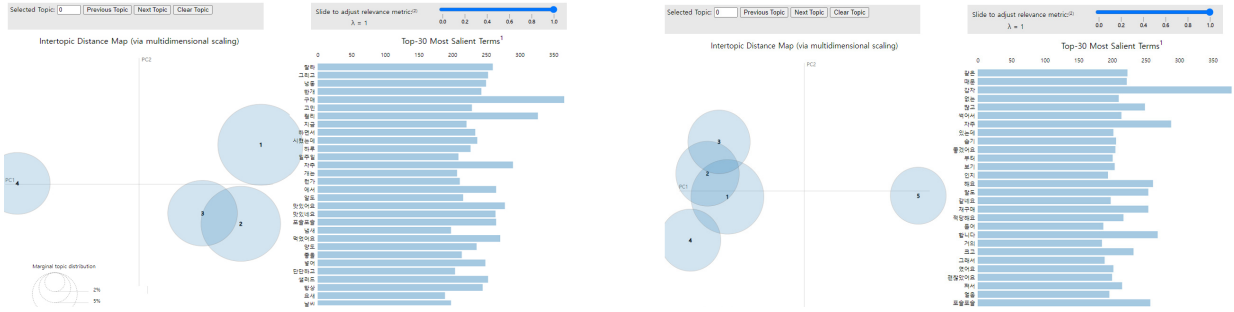
4.3 최종모델링 결과

CTM 모델링을 시각화한 그림은 Fig. 8과 같으며, 순차적으로 2~5개로 토픽 개수를 늘려가며 시각화를 진행하였다[26]. 토픽이 2개로 나누어진 경우에는 두 개 토픽 내 여러 주제가 혼합되어 연관성이 낮아 분류의 의미를 찾을 수가 없었다. 반대로 토픽 수가 4개 이상부터는 서로 겹치는 영역이 현저히 증가하는 것을 시각적으로 확인할 수 있었다. 즉, 4개 이상 토픽 모델링에서는 토픽별로 중첩되어 판별타당도가 낮아지는 양상을 보였다. 그 결과 본 신선식품 구매후기 데이터 분석 연구의 설명가능한 최적 토픽 개수는 3개로 판단할 수 있었다.



(a) (b)

Fig. 8. CTM Topic modeling results: (a) Topic 2, (b) Topic 3, (b) Topic 4 and (b) Topic 5



(c) (d)

Fig. 8. CTM Topic modeling results: (a) Topic 2, (b) Topic 3, (b) Topic 4 and (b) Topic 5 (continued)

CTM은 LDA에 비해 상대적으로 재현성이 높긴 하지만, CTM 또한 모델을 학습시킬 때마다 약간의 차이가 발생했다. 따라서 결과 앙상블(ensemble)기법을 통해 재현성을 더 높이고자 했다[25]. 이를 토대로 토픽별로 가장 많이 등장한 상위 5개 단어를 선택한 결과는 아래 Table 2와 같다. 이를 토픽별로 자세히 살펴보면, Topic. 1(전체 토큰 중 42.4%)은 감자의 배송 및 포장 상태에 관련된 토픽, Topic. 2 (전체 중 32.4%)는 감자의 특성과 상태, Topic. 3(전체 중 25.2%)은 감자의 요리법과 관련된 토픽이 나왔음을 확인할 수 있었다[27]. 이에 해당 구매후기 데이터 토픽 그룹별로 가장 적합하게 각각 배송 및 포장, 신선식품의 상품의 특징 및 상태, 그리고 구매 목적 및 요리법으로 크게 3개의 명칭 주제로 분류할 수 있었다[28-29].

Table 2. Representative keywords by Topic

Topic No.	Classification	Top First Word	Top Second Word	Top Third Word	Top Fourth Word	Top Fifth Word
Topic.1 (42.4%)	Delivery and Packaging (배송 및 포장)	Moderate (적당, 12)	Fresh (싱싱, 9)	New (신선, 7)	Clean (깨끗 5)	Convinient (편해요, 4)
Topic.2 (32.4%)	Product Features and Conditions (상품 특징 및 상태)	Single Size of Potatoes (알크기, 10)	Sparsely (포슬포슬, 8)	Solid (단단, 7)	Big-size (크고, 6)	Just-Harvested (햇수확한, 4)
Topic.3 (25.2%)	Purpose of Purchase and Recipes (구매 목적 및 요리법)	Stir-fried (볶음, 8)	Curry (카레, 7)	Powder (가루, 6)	Salad (샐러드, 4)	Side-dish (반찬, 2)

5. 결론 및 제언

5.1 연구결과 요약

전 세계적으로 온라인 식료품 구매율이 증가함에 따라 소비자들의 이용후기가 구매행동에 큰 영향을 끼치고 있다. 이러한 과정에서 구매후기를 심층적으로 분석하고 고객의 관점에서 가치 있는 정보로 제공하여 고객경험을 향상하는 일이 중요해졌다[27]. 이에 반해 기존 후기 연구에서는 '좋아요' '만족' '추천'과 같은 범용적인 단어를 추출한 해석의 한계가 있어, 실제로 상품군의 특징과 구매동기를 파악하기가 어려웠다. 이에 본 연구에서는 고객후기 데이터를 토픽 모델링 분석하여 연관도가 높은 키워드를 중심으로 소비자들의 구매여정에 접목 후 실질적인 가치를 제공하고자 했다[28].

이에 국내 대표 식품전문몰 베스트 상품에 게시된 후기를 수집하여 BERTopic, LDA, CTM 등 3종의 토픽 모델링을 적용하여 성능을 비교하였다. BERT의 경우, 텍스트 데이터에서 의미론적으로 관련된 문서들을 클러스터링하여 토픽을 추출하는 방법이

나 여러 토픽을 선택했을 때 한 토픽에 단어들이 크게 몰리는 단점이 있었다. 또한 단어의 출현 빈도와 토픽의 분포를 결합하여 각 문서의 토픽 분포를 추정하는 LDA의 경우, 무작위로 토픽이 할당되므로 분석 결과의 재현성이 떨어져 본 연구의 최종모형에 적합하지 않았다. 그 결과 가장 성능이 좋은 CTM을 선정하였고, 79,616건의 후기 데이터 중 심층적인 특징을 탐구할 수 있는 유효한 3가지 토픽인 배송 상태, 상품의 특성, 그리고 요리법 관련 주제를 추출할 수 있었다.

5.2 시사점 및 후속연구 제언

온라인 식품전문몰은 본 연구에서 후기 데이터 분석 결과를 활용한 서비스를 통해 고객들에게 새로운 가치를 제공해 줄 수 있을 것이다. 다시 말해 식료품 구매 전 서비스 내 구매후기 데이터를 기반으로 한 고객 관점의 이용 정보를 구체적으로 제공하여 신뢰할 수 있는 구매 행동으로 연계할 수 있다. 이는 소비자 경험 측면에서 해당 서비스 고객들에게 필요한 후기 정보는 키워드를 통해 직관적으로 제공하여 고객들의 시간과 노력을 줄이는 효용도 제공할 것이다. 카테고리별로 구분된 후기 키워드를 제공한다면 더 많은 고객의 구매 행동을 끌어낼 수 있을 것으로 기대된다. 궁극적으로 의미있고 가치있는 상품평의 전달은 플랫폼과 브랜드 신뢰도를 제고하고, 고객경험을 향상할 수 있다. 나아가 고객들의 우수한 사용자경험은 지속적으로 방문하고 탐색하는 트래픽을 증가시키며, 이는 매출과 수익성이 증가하는 선순환 가치를 구축하는 토대가 될 것이다[28]. 다음으로 데이터 분석 활용의 측면에서, 온라인 식품전문몰 플랫폼은 소비자들의 생생한 후기 데이터를 활용하여 상품 자체의 부족한 점을 보완할 수 있을 것이다. 다양한 상품의 후기 데이터를 수집하여 입체적으로 분석한다면, 온라인 식료품 구매에서의 고객들의 불만족을 줄이고 구매이탈을 막을 수 있을 것이다[29]. 또한 보다 정확하게 고객들에게 필요한 키워드를 추출하여 제공하는 것은 물론 제품기획 및 개발, 개선 관점에서 큰 도움이 될 것이다. 최근 거대언어모델(LLM) 기반 OpenAI ChatGPT, Google Bard & Gemini, MS Bing, WRTN(뤼튼) 등 생성형 AI의 API를 활용하여 후기 작성에 있어 추천 문장 등을 적극적으로 제공한다면, 양질의 고객경험을 묘사하고 작성할 수 있을 것으로 기대된다[30].

마지막으로 더 나은 후속 연구를 위해 다음과 같은 제언을 하고자 한다. 먼저 현재 모델에서는 후기 데이터 분석을 통해 카테고리별로 추출된 키워드에 대하여 소비자들에게 도움이 되는 토픽(주제)으로 명확하게 제시 및 전달되는지에 대하여 모호함이 있다. 유효 키워드의 모호함을 개선하기 위해 사용자 조사와 함께 웹 시연을 통해 얻은 인사이트를 기반으로 소비자들에게 도움이 되는 키워드 위주로 사용자화면(인터페이스)을 개선할 수 있을 것이다. 아울러 본 연구의 시공간적, 물리적 한계로 인해 대표 상품의 구매후기 데이터 활용으로 한정되었지만, 이러한 분석 프레임워크를 향후 카테고리별로 확장 적용한다면 새로운 시사점을 도출해 낼 수 있을 것이다. 과일, 견과, 양곡, 반찬은 물론 수산물, 나아가 건강 기능성/유기농/비건(Vegan)식품, 반려동물용 식품 등 종류와 형태에 따라 다채로운 특성을 보일 것이다. 앞으로 온라인 식품 전문몰 플랫폼에서 고객들의 후기 분석 등 가치있는 정보를 제공하는 것은 장기적으로 고객경험과 고객가치를 향상하는 데 큰 영향을 줄 것으로 보인다.

References

[1] eMarketer, “2022 Total Retail Sales Worldwide”, www.insiderintelligence.com (accessed December 31, 2023).

[2] KREI, “Food Consumption Behavior Survey Report 2024”, pp. 5-19, www.krei.re.kr (accessed March 23, 2024).

[3] Statistics Korea, “2023 Annual Online Shopping Trends”, pp. 1-15, www.kostat.go.kr (accessed March 23, 2024).

[4] OpenSurvey, “2023 Online Grocery Buying Trends Report”, pp. 3-17, www.opensurvey.com (accessed February 3, 2024).

[5] H.J. Lee, S.Y. Kwon, and D.H. Min, “The Empirical Research on the User Satisfaction of Mobile Grocery Shopping Customer Journey”, Journal of Information Technology Applications & Management(JITAM), Vol. 28, No. 4, pp. 59-78, 2021.

[6] Consumers Union of Korea, “Survey Results on the Status and Aware of Consumption”, www.cuk.or.kr (accessed March 7, 2024).

[7] R. Rietsche, D. Frei, E. Stoeckli, and M. Söllne, “Not All Reviews are Equal - a Literature Review on Online Review Helpfulness”,

In Proceedings of the 27th European Conference on Information Systems (ECIS), Stockholm & Uppsala, Sweden, pp. 1-18, 2019.

[8] K. Zhao., A. C. Stylianou, and Y. Zheng, "Sources and impacts of social influence from online anonymous user reviews", Information & Management, Vol. 55, No. 11, pp. 16-30, 2018.

[9] S. M. Mudambi, and D. Schuff, "What makes a helpful online review? A study of customer reviews on Amazon.com", MIS Quarterly, Vol. 34, No. 1, pp. 185-200, 2010.

[10] M. Li, L. Huang, C. Tan, and K. Wei, "Helpfulness of online product reviews as seen by consumers: Source and content features", International Journal of Electronic Commerce(IJEC), Vol. 17, No. 4, pp. 101-136, 2013.

[11] H. G. Lee, and H. Gwak, "A Study on the Factors Influencing the Effectiveness of Online Consumer Reviews and Informatization Policy", Korea Intelligence Society Agency(KISA), Vol. 20, No. 3, pp. 3-17, 2013.

[12] K. Saleh, "The Importance of Online Customer Reviews", retrieved from www.invespcro.com/blog/the-importance-of-online-customer-reviews-infographic (accessed June 10 2023)

[13] S. W. Ahn, "Effects of Customer Rating and Review on Purchase Intention of Experience Product: Moderating Effects of Preference Similarity and Preference Differentiation", Korean Society of Consumer Studies, Vol. 33, No. 4, pp. 1-25, 2022.

[14] J. Han, and J. S. Kim, "The Effect of Review Attributes on Brand Attitude, Purchase Decision and e-WOM Intention in Online Shopping Mall", The Society of Digital Policy & Management(SDPM), Vol. 19, No. 7, pp. 113-127, 2021.

[15] F. Zhu, and X. Zhang, "Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics", Journal of Marketing, Vol. 74, No. 2, pp. 133-148, 2010.

[16] W. Chu, and M. Roh, "Exploring the Role of Preference Heterogeneity and Causal Attribution in Online Rating Dynamics", Asia Marketing Journal, Vol. 15, No. 4, pp. 61-101, 2014.

[17] J. Kim, and S. J. Lennon, "Effects of reputation and website quality on online consumers' emotion, perceived risk and purchase intention", Journal of Research in Interactive Marketing(JRIM), Vol. 7, No. 1, pp. 33-56, 2013.

[18] Z. Jiang, and I. Benbasat, "Virtual product experience: Effects of visual and functional control of products on perceived diagnosticity and flow in electronic shopping", Journal of Management Information Systems, Vol. 21, No. 3, pp. 111-147, 2004.

[19] P. A. Pavlou, H. Liang, and Y. Xue, "Understanding and mitigating uncertainty in online exchange relationships: A principal-agent perspective", MIS Quarterly, Vol. 31, No. 1, pp. 105-136, 2007.

[20] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure", arXiv eprint, arXiv:2203.05794, https://arxiv.org/abs/2203.05794, Vol. 1, 2022.

[21] F. Bianchi, S. Terragni, and D. Hovy, "Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence", Association for Computational Linguistics, Vol. 2, pp. 759-766, 2021.

[22] J. Devlin, M. Y. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding", arXiv preprint, arXiv:1810.04805. https://arxiv.org/abs/1810.04805, Vol. 1, 2018.

[23] D. M. Blei, A. Y. Ng, and M. I. Jordan. "Latent Dirichlet Allocation", Journal of Machine Learning Research, Vol. 1, 2003.

[24] FAO STAT, https://faostat.fao.org (accessed December 21, 2023)

[25] J. Chuang, C. D. Manning, and J. Heer, "Termite: Visualization Techniques for Assessing Textual Topic Models", Proceedings of the International Working Conference on Advanced Visual Interfaces, pp. 74-77, 2012.

[26] C. Sievert, and K. E. Shirley, "LDAvis: A method for visualizing and interpreting topics", Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, pp. 63-70, 2014.

[27] Y. G. Park, D. H. Min, and H. J. Lee, "The Effect of Live Broadcast of Fresh Food on Customer's Purchasing Intention", The Journal of Industrial Distribution & Business, Vol. 14, No.9, pp. 31-39, 2023.

[28] J. H. Kim, Y. B. Ko, J. H. Choi, and H. J. Lee, "Research on the Design of a Deep Learning-Based Automatic Web Page Generation System", Journal of The Korea Society of Computer and Information (JKSCI), Vol. 29, No. 2, pp. 21-30, 2024.

[29] H. Lee, Y. G. Park, and D. H. Min, "Analysis of Factors Affecting the Continuance Intention to Use Mobile Grocery Shopping", The Journal of Information Systems, Vol. 29, No. 2, pp. 95-110, 2020.

[30] S. H. Park, Y. E. Lee, and H. J. Lee, "Research on Enhancing Customer Experience through AI-Supported Review Generation", Transactions of The Korean Institute of Electrical Engineers(TKIEE), Vol. 73, No. 2, pp. 334-342, 2024.