Journal of the Korean Institute of Industrial Engineers
Vol. 51, No. 6, pp. 489-501, December 2025.
ISSN 1225-0988 | EISSN 2234-6457
https://doi.org/10.7232/JKIIE.2025.51.6.489

텍스트 마이닝 기반 모바일 금융앱 UX 문제 진단 및 개선안 제안

박가율[1] · 정명철[1] · 모승민[2†]

[1]아주대학교 산업공학과 / [2]오산대학교 안전보건관리과

Text Mining-Based Diagnosis and Improvement of UX Issues in Mobile Financial Applications

Ga-yul Park[1] · Myung-Chul Jung[1] · Seung-Min Mo[2]

[1]Department of Industrial Engineering, Ajou University
[2]Department of Occupational Safety and Health, Osan University

This study explores user experience issues in mobile financial applications by analyzing user reviews and identifying areas of satisfaction and dissatisfaction. Large-scale sentiment classification was performed using a model trained with practically labeled data, enabling efficient processing of over 130,000 user reviews. The analysis revealed key strengths such as intuitive navigation, smooth transactions, and clear information delivery, which suggest opportunities for improving personalization, UI consistency, and usability across devices. Conversely, common issues included authentication failures, unintuitive menus, and server instability, highlighting the need for simplified login processes, better interface structure, and technical reliability. Based on these insights, the study proposes actionable strategies to enhance user experience and outlines practical implications for data labeling and review analysis in UX diagnostics.

Keywords: Text Mining, UX, Sentiment Analysis, Topic Modeling, Mobile Apps

1. 서 론

비대면 금융 서비스의 확산과 디지털 전환의 가속화(Hong and Pan, 2020)는 모바일 금융앱이 사용자 일상에 깊숙이 자리 잡는 계기가 되었다. 은행, 카드사, 핀테크 기업 등의 금융기관이 모바일 플랫폼을 통해 계좌 조회, 송금, 투자, 대출 등 다양한 금융 서비스를 제공하고 있기 때문에 사용자는 대부분의 금융 업무를 모바일로 간편하게 처리하고 있다(Rho, 2021). 이러한 변화 속에서 금융앱의 사용자 경험(User Experience, UX)은 단순한 기능적 편의성을 넘어 브랜드 신뢰도와 고객 만족도를 좌우하는 핵심 경쟁 요소로 부상하였다(Koh and Pan, 2021). 특히 금융앱은 복잡한 정보 구조와 높은 보안성 요구 등으로 인해 일반 모바일 앱보다 정교한 UX 설계가 요구되며, 금융기관에서도 UX 향상을 위한 다양한 노력이 이루어지고 있다.

기존의 UX 연구는 사용자 모집을 통한 설문조사와 인터뷰 방식이 중심이었으나, 사전 구성된 질문이나 환경에 따라 실제 사용자의 복잡하고 세밀한 사용 경험을 반영하는 데 한계가 있다(Kim *et al.*, 2025). 이에 따라 최근에는 사용자 리뷰 기반의 감성분석을 통해 UX 개선 요소를 도출하려는 연구가 증가하는 추세이다. 사용자 리뷰는 실사용 경험을 바탕으로 작성된 비정형 텍스트 데이터와 평점 기반의 정형 데이터로(Jin and Lee, 2022), UX 개선 요소를 발굴하는 데 유용한 자료가 된다. 다만, 사용자들이 남기는 대량의 리뷰를 수작업으로 분석하는 것은 많은 시간과 비용을 지불해야 하기 때문에(Son *et*

† 연락저자 : 모승민 교수, 18119 경기도 오산시 청학로 45 오산대학교 안전보건관리과, Tel : 031-370-2701, Fax : 031-370-2709, E-mail : smmo@osan.ac.kr
2025년 7월 18일 접수; 2025년 9월 3일; 2025년 9월 13일 수정본 접수; 2025년 9월 15일 게재 확정.

al., 2021), 국내에서는 한국어 특화 딥러닝 언어 모델을 통한 사용자 리뷰 감성분석 연구가 활발히 진행되고 있다.

대부분의 기존 연구에서는 감성분석 모델에 활용될 학습 데이터 구축을 위해 리뷰 평점을 기반으로 긍정과 부정을 나누는 평점 기반 라벨링 방식을 사용해왔다(Park and Bae, 2024; Lee *et al.*, 2023; Li and Wu, 2024). 이 방식은 대규모 데이터 구축에 용이하다는 점에서 널리 활용되고 있지만, 실제 텍스트의 감정과 평점 간의 불일치로 인해 모델 분류 정확도 저하로 이어질 수 있다는 문제점이 제기되고 있다(Kim and Han, 2022). 특히 많은 선행연구에서 평점 기반 라벨링 방식을 채택하고 있으나, 해당 방식을 선택한 이유나 그 유효성을 평가한 근거는 구체적으로 제시되지 않는 경우가 대부분이다. 반면, 연구자가 직접 리뷰를 읽고 긍정과 부정을 분류하는 수작업 라벨링 방식은 정밀한 데이터 구축이 가능하지만, 연구자의 부주의나 잠재적 속성이 영향을 주어 예측 편향이 일어날 수 있다(Park, 2022). 또한 수작업 라벨링은 시간 및 비용면에서 비효율적이며 인간에 의한 데이터 품질 저하도 배제할 수 없기 때문에(Kang *et al.*, 2024) 대규모 데이터 구축의 실용성이 떨어진다. 따라서 학습 데이터 구축 방식이 딥러닝 언어 모델의 감성 분류 성능에 어떠한 영향을 미치는지에 대한 비교 및 검증은 학습 데이터 구축 전략을 수립하고, 분석 정확도와 효율성 간의 균형을 판단하는 데 중요한 의미를 갖는다. 본 연구는 이러한 필요에 따라 수작업 라벨링 방식과 평점 기반 라벨링 방식 간의 성능을 정량적으로 비교하고, 그 결과를 바탕으로 딥러닝 언어 모델의 학습 데이터 라벨링 전략 수립에 실질적인 근거를 제시하고자 한다. 또한 감성분석은 리뷰에 나타난 긍정과 부정의 감성을 효율적으로 분류하는 데에 유용하지만, 단순 분류는 사용자의 구체적인 만족 및 불만족 요소를 분석하는 데 제한적일 수 있다(Hyun *et al.*, 2019). 따라서 리뷰의 핵심 주제를 심층적으로 파악하기 위해서는 감성분석과 더불어 토픽 모델링과 같은 추가적인 분석 기법이 요구된다. 토픽 모델링은 대량의 텍스트 속에서 중요한 의미를 지닌 토픽을 도출해주는 분석 기법으로(Lee, 2024), 반복적으로 언급되는 주제와 키워드를 추출하여 감성분석의 한계를 보완하는 데 효과적인 기법으로 활용될 수 있다. 마지막으로, 기존 연구들은 대부분 단일 금융앱 혹은 2개의 금융앱에 대한 비교 분석 등 소수의 분석 대상을 다루는 경우가 많아, 도출된 UX 개선안이 개별 앱에 국한되는 경향이 있다. 이로 인해 결과의 일반화 가능성이 제한되며, 금융 분야 전반에 적용 가능한 UX 개선안을 도출하는 데 한계가 존재한다. 따라서 다양한 유형의 금융앱을 포괄한 사용자 리뷰를 기반으로, 공통적으로 나타나는 UX 요소를 함께 분석하려는 시도가 필요하다.

이에 본 연구는 모바일 금융앱 분야에서의 사용자 리뷰를 수집하고, 학습 데이터 구축 방식에 따른 딥러닝 언어 모델의 감성분석 성능을 정량적으로 비교함으로써, 학습 데이터 구축 방식이 모델 성능에 미치는 영향을 확인하고자 한다. 이후 성능이 더 우수한 라벨링 방식을 통해 미세조정된 딥러닝 언어 모델을 활용하여 감성분석을 수행하고, 토픽 모델링을 통해 도출된 사용자 경험의 긍정 및 부정 요소를 피터 모빌의 허니콤 모델을 바탕으로 긍정 요소의 강화 방안과 부정 요소의 개선 방안을 제시하고자 한다.

2. 선행연구

2.1 UX 연구

기존 UX 연구는 사용자 실험을 병행한 설문조사 및 인터뷰 방식이 중심이었다. Han and Lim(2014)는 지문인식 방식이 금융앱의 공인인증서 보안 방식을 대체할 수 있다는 가정 하에 금융앱 서비스에 대한 UX 평가 및 분석을 실시하였다. 이에 20~50대의 금융앱 사용자 대상 설문조사를 통해 UX 평가 기준을 도출하고, 4가지 유형의 지문인식 방식을 구현한 프로토타입으로 과업을 수행하고 리커트 척도의 설문조사와 인터뷰를 실시하였다. 결론적으로 사용자들은 스마트폰 환경에 좀 더 적합하고 편리하여 보안성에 긍정적인 영향을 주는 지문인식 방식을 선호한다는 결과가 도출되었다. Kim and Kim(2020)은 2개의 간편 결제 앱을 대상으로 세대 간의 사용자 경험을 알아보고자 허니콤 모델 기반 설문조사와 3가지 과업이 포함된 심층 인터뷰를 통해 4가지 개선 방안을 제시하였다. Jo and Chang(2020)은 고령화 사회로 인한 실버세대의 모바일 금융앱 사용 편의성과 접근성을 향상시키기 위해 국내 4개 금융앱을 대상으로 UI 구성요소를 비교하고, 60대 대상 설문조사와 심층 인터뷰를 통해 사용성 평가를 진행하였다. 이에 대중적인 아이콘과 버튼의 사용이 인지적 용이성에 도움을 주며, 버튼의 색상으로 중요 버튼을 강조하여 주목도를 높이는 것을 권장하는 등 7개의 개선안을 제시하였다. Jung and Yeoun(2022)는 금융 마이데이터 서비스의 사용자 경험 향상을 목적으로 심층 인터뷰를 통해 영향 요인을 도출하고, 주 사용 메뉴를 선정 후 분석하였다. 주 사용 메뉴의 제공 방식, 화면 구성 등을 분석하여 핀테크와 은행의 차이를 비교하고, 설문조사를 통해 메뉴별 요인을 도출하여 플랫폼 선호도를 알아보았다. 결과적으로 6개의 영향 요인을 제시하였으며, 가장 유의미한 요인은 직관성이며, 편의성과 신뢰성 또한 주요 요인이라고 제시하였다. 이 같은 설문조사 및 심층 인터뷰 방식은 연구 목적에 따라 다양한 측정이 가능한 장점이 있으나, 설문의 형태와 용어 등에 따른 측정 오류의 발생 가능성이 있다(Yun and Choi, 2021). 또한 응답자가 설문과 관계없는 부적절한 응답을 제공할 경우 연구 결과가 왜곡될 위험성이 존재한다(Park *et al.*, 2020).

2.2 감성분석

감성분석이란 데이터 마이닝 및 자연어 처리의 하위 분야로, 비정형적인 텍스트에서 연구 대상에 대한 긍부정 및 중립

의 판별을 분석한다(Yun, 2025). 오피니언 마이닝이라고도 불리며 리뷰, 설문 응답, 소셜 미디어 등의 분석에 활용된다(Lee and Jang, 2024). 이처럼 다양한 분야의 감성분석을 위해 BERT, ELECTRA 등의 딥러닝 언어 모델을 통해 텍스트의 주요 뉘앙스를 파악하여 분류하는 딥러닝 기반 감성분석 연구가 활발히 수행되고 있다(Kim and Cho, 2025). Lee *et al.*(2023)은 앱 개발자들이 리뷰의 감성을 빠르고 효과적으로 파악할 수 있는 리뷰 감성분석 앱을 구현하고자 하였다. 이에 구현할 앱에 필요한 언어 모델의 선정을 위해 한국어 특화 언어 모델인 KoBERT, KoGPT-2, KoBART의 성능을 비교 분석하여 가장 좋은 성능을 보인 KoBART 모델로 리뷰 감성분석 앱을 구현하였다. Jeon and Noh(2025)는 온라인 패션 제품의 고객 경험 인사이트 도출을 위해 KoELECTRA를 활용한 리뷰 감성분석을 통해 소비자 중심의 중요 속성을 도출하였다. Jung *et al.*(2021)은 토픽 모델링과 시계열 이상치 탐지를 통해 리뷰의 이상치를 탐지하여 주제를 분류하고자 하였다. 이에 TEANAPS 패키지에서 제공하는 KoBERT를 통해 이상치로 분류된 리뷰의 감성분석을 수행하고 평균 부정 리뷰 비율을 도출하였다.

기존 연구에서 활용된 여러 한국어 특화 언어 모델 중 KoBERT(Korean Bidirectional Encoder Representations from Transformers) 모델은 SKT가 뉴스 및 위키피디아의 한국어 문장을 학습시켜 개발한 한국어 특화 BERT 모델로(Park, 2025), 한국어의 복잡한 문법 구조와 표현 방식에 특화되어 한국어를 효과적으로 분석할 수 있다는 강점을 가지고 있다(Kim *et al.*, 2024a). 이에 본 연구에서도 KoBERT 모델을 채택하여 연구를 수행하였다. 그러나 사전학습 언어 모델인 KoBERT는 일반적인 문장들을 학습하였기 때문에 전문적인 분야에서 제한적으로 사용되는 용어는 정확하게 이해하지 못하는 한계가 존재한다(Han *et al.*, 2022). 따라서 본 연구에서는 모바일 금융앱 분야에서의 적용을 위해 금융 분야의 용어를 추가로 학습하기로 했다.

2.3 토픽 모델링

토픽 모델링은 문장의 맥락과 관련된 단서들을 통해 유사한 의미를 가진 단어들을 클러스터링하는 방식으로 주제를 추측하는 알고리즘이다(Kang *et al.*, 2013). 어떤 주제들의 집합이라고 가정된 구성 단어들을 확률적으로 계산한 결과값을 통해 토픽 주제어들의 집합을 추출하는 것이다(Bae *et al.*, 2014). 이를 통해 대규모 텍스트 데이터의 분석 작업을 자동화 분석할 수 있다(Lee and Kim, 2023). 기존에는 LDA(Latent Dirichlet Allocation)가 주로 사용되었는데, 가장 핵심 주제를 파악할 수 있다는 장점이 있지만, 최적의 토픽 개수를 결정하는 과정이 필요하고, 단어의 의미를 반영하지 못한다는 단점이 있다(Lee and Kim, 2024). 따라서 최근에는 LDA의 한계점을 극복하기 위한 BERTopic이 활용되고 있다. BERTopic은 LDA와는 다른 임베딩 방식을 통해 단어의 의미를 반영하고, 최적의 토픽 개수를 자동으로 결정하는 등 편의성과 객관성이 높다(Jo and Kim, 2025). BERTopic은 사전학습된 언어 모델을 통해 입력 데이터를 임베딩하고, 밀도 기반 클러스터링 방법인 UMAP(Uniform Manifold Approximation and Projection)을 거친 후 HDBSCAN(Hierarchical Density-Based Spatial Clustering of Applications with Noise) 알고리즘으로 토픽을 군집화하여 class-based TF-IDF(Term Frequency-Inverse Document Frequency)를 통해 토픽의 일관성 있는 표현을 추출하는 것이다(Lee *et al.*, 2024). 이에 따라 본 연구에서도 LDA의 한계점을 극복할 수 있는 BERTopic을 활용하였다.

2.4 허니콤 모델

허니콤 모델은 피터 모빌(Peter Morvile)에 의해 제안되었으며, User Experience Honeycomb이라 불리는 벌집 모양으로 플랫폼의 주요 기능과 속성을 토대로 사용자의 경험 요소를 평가하는 툴로 사용되고 있다(Jung, 2024). 7가지의 사용성 원칙을 기반으로 평가하는데, 유용성(Usefulness)은 서비스가 사용자에게 실제적인 이익과 도움을 제공하는 정도를 의미하며, 사용성(Usability)은 서비스 이용 과정에서 느끼는 편리함과 효율성을 나타낸다. 매력성(Desirability)은 시각적 디자인과 전반적인 사용 경험이 얼마나 매력적이고 호감 있게 인식되는지를 뜻하며, 신뢰성(Credibility)은 데이터 보안, 안정성, 전문성 등에 대한 사용자의 신뢰 수준을 반영한다. 접근성(Accessibility)은 사용자가 서비스에 얼마나 손쉽게 접근할 수 있는지를 평가하며, 가치성(Value)은 서비스에서 제공되는 전반적인 가치를 사용자가 어떻게 인식하는지를 보여준다. 마지막으로 검색성(Findability)은 원하는 기능이나 정보를 얼마나 쉽게 탐색하고 찾아낼 수 있는지를 의미한다(La *et al.*, 2024). 허니콤 모델은 평가 형식에 의해서 감정(Inspection)을 위해 활용할 수 있으며, 평가 성격 측면에서는 검증(Validation), 측정(Assessment), 탐색(Exploratory), 비교(Comparison)를 위해 다양하게 활용된다(Si *et al.*, 2019). 본 연구에서도 토픽 모델링의 결과로 도출된 긍정과 부정 키워드 해석에서의 자의적 해석을 최소화하고자 허니콤 모델에 대응시켜 객관성을 확보하고자 하였다.

3. 연구방법

본 연구는 모바일 금융앱 리뷰 데이터를 기반으로 수작업 라벨링과 평점 기반 라벨링 방식에 따른 KoBERT 감성분석 모델의 성능을 비교하고, 성능이 우수한 모델을 선정하였다. 선정된 모델로 금융앱 리뷰의 감성분석을 수행하고 BERTopic 기반 토픽 모델링을 통해 긍정 및 부정 리뷰의 주요 UX 요소를 도출하여 피터 모빌의 허니콤 모델을 토대로 긍정 리뷰에서의

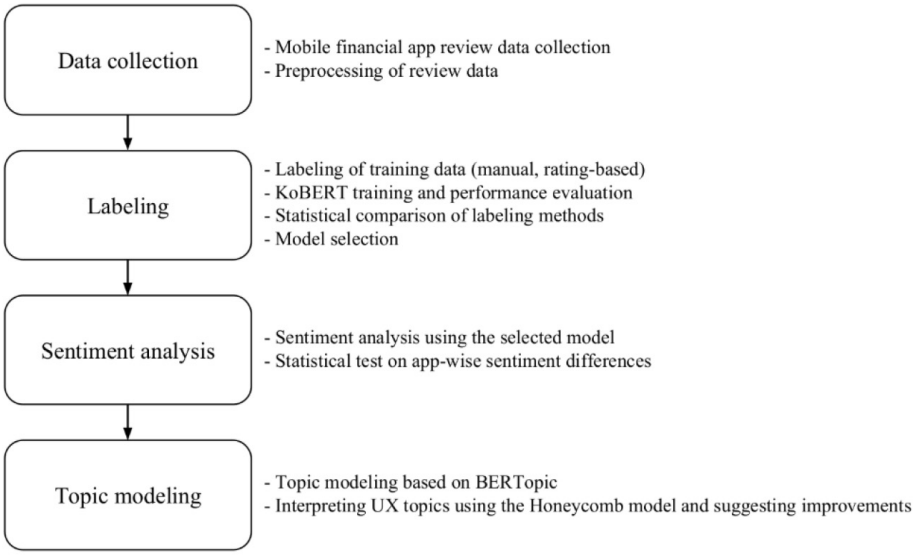

Figure 1. Flowchart of the Analysis Process

강화 방안과 부정 리뷰에서의 개선 방안을 제시하였다. 본 연구의 절차는 <Figure 1>과 같다.

3.1 데이터 수집 및 전처리

금융앱 선정은 2025년 3월 컨슈머인사이트가 전국 20~69세 금융소비자 2,083명을 대상으로 실시한 금융앱 확보고객 순위와, 한국기업평판연구소가 2025년 3월 8일부터 4월 8일까지 수집한 지방은행 브랜드 빅데이터(9,200,705건)를 기반으로 발표한 브랜드 평판 순위를 참고하여 최종 25개의 금융앱을 선정하였다. 선정한 25개의 금융앱을 금융권역별로 분류한 표는 <Table 1>과 같다. 각 금융앱의 리뷰는 오픈 소스 프로그램인 Python 3.11.9와 웹 크롤링 라이브러리인 google-play-scraper을 통해 구글 플레이스토어에서 수집되었다. 각 금융앱의 리뷰는 최초 작성 시점부터 2025년 4월 20일까지 총 574,564건이 수집되었다. 수집된 리뷰들은 중복 내용 제거, 5자 미만 리뷰 삭제, 이모티콘 및 단순 표현 삭제, 반복 문자 제거, Selenium 패키지를 활용한 맞춤법 및 띄어쓰기 교정의 전처리 과정을 거쳐 최종적으로 448,918건의 리뷰가 확보되었다.

3.2 데이터 구성

전처리된 리뷰는 학습용 데이터와 분석용 데이터로 구분하였다. 학습용 데이터는 금융분야에서 사용되는 용어 및 자주 쓰는 표현들을 모델에 추가적으로 학습시키기 위하여 활용되는 데이터로, 리뷰 최초 작성 시점부터 2021년까지의 리뷰 314,358건으로 구성하였다. 딥러닝 언어 모델의 성능을 높이기 위해서는 충분한 양과 품질의 학습용 데이터가 필요한데

Table 1. List of 25 financial apps categorized by financial sector

No.	Financial sector	Financial app	No.	Financial sector	Financial app
1	Commercial banks	A	14	Fintech services	N
2		B	15		O
3		C	16		P
4		D	17	Credit card companies	Q
5		E	18		R
6		F	19		S
7	Regional banks	G	20		T
8		H	21		U
9		I	22		V
10		J	23		W
11		K	24		X
12	Internet-only banks	L	25		Y
13		M			

(Lee and Lee, 2022), 모바일 앱은 업데이트가 빈번히 이루어져 (Kim *et al.*, 2014) 과거의 리뷰는 최근 UX 분석에는 적합하지 않다고 판단하였다. 따라서 과거 리뷰를 학습용 데이터로 사용해 모델이 금융 리뷰의 문맥을 폭넓게 이해하여 분석 성능을 강화할 수 있도록 하였다. 또한 앱별 리뷰 수의 분포가 불균형한 점을 고려하여, 전체 금융앱 리뷰 수의 평균값인 12,574건을 학습용 데이터의 최소 기준선으로 설정하였다. 이는 리뷰 수가 적은 앱을 포함할 경우 특정 앱의 특성에 모델이 과도하게 적합(Overfitting)되거나, 전체 데이터 분포를 대표하지 못하는 편향 학습이 발생할 수 있기 때문이다(Cho and Oh, 2022). 따라서 최종적으로 학습용 데이터로 활용된 8개 금융앱은 N(60,118건), A(54,398건), V(26,899건), T(25,694건), B(19,983건), Q(19,885건), S(14,372건), F(13,958건)이며, 총 235,307건이다. 단, 금융앱 J는 2023년 3월 출시되어 2021년까지의 리뷰가 존재하지 않아 분석 데이터에서만 활용되었다.

분석용 데이터는 실제 감성분석 및 토픽 모델링에 활용되는 데이터로, 2022년부터 2025년 4월 20일까지의 최근 3년 리뷰 134,560건으로 구성하였다. 분석용 데이터는 전체 25개 금융앱의 리뷰를 모두 포함하였는데, 이는 서론에서 언급한 바와 같이 기존 연구가 소수 범위의 앱에 한정될 경우 UX 개선 요소가 특정 앱에 국한되어 결과의 일반화가 제한될 수 있기 때문이다. 따라서 분석용 데이터는 금융앱 전반의 공통적 사용자 경험을 반영하도록 전체 금융앱의 최근 3년 리뷰를 모두 포함하였다.

최종적으로 학습용 데이터와 분석용 데이터에 활용된 앱과 리뷰 수를 제시한 표는 <Table 2>와 같으며, <Figure 2>은 전반적인 데이터 정제 과정을 개략적으로 나타낸 것이다.

Table 2. Number of Reviews in the Preprocessed Training and Analysis Datasets by Financial App

Dataset type	Apps used	Number of reviews
Training set	8 apps (A, B, F, N, Q, S, T, V)	235,307
Analysis set	25 apps	134,560

3.3 라벨링

학습용 데이터 235,307건 중 일부를 무작위 샘플링하여 수작업으로 감정 라벨링을 수행하였다. 전체 데이터를 수작업으로 라벨링하는 것이 현실적으로 어려운 상황에서, 편향이 없고 신뢰도 있는 학습용 데이터 구성을 위해 무작위 샘플링 기법을 사용하였다(Shin *et al.*, 2023). 이 과정에서 긍정 리뷰의 수가 상대적으로 적게 확보되어 클래스 불균형이 발생하였으며, 학습 데이터가 많은 클래스에 모델이 편향되어 성능 저하 문제로 이어질 수 있었다(Kim and Chung, 2024). 이를 보완하기 위해 부정 리뷰를 언더샘플링하여 최종적으로 55,738건(긍정 27,869건, 부정 27,869건)의 학습용 데이터를 구축하였다. 수작업 라벨링은 연구자를 포함한 3인의 전문가가 직접 리뷰 본문을 판독하여 수행하였으며, 앱 사용성에 대한 명확한 감정 및 평가가 드러난 리뷰들을 긍정 또는 부정으로 분류하였다. 반면 단순 요청, 문의, 중립적 표현, 감정이 모호한 리뷰 등은 기준에 따라 제외하였다. 라벨링 과정 중 이견이 발생한 경우에는 해당 표현의 사전적 의미를 검토하여 긍정과 부정을 결정하거나 제외하였다.

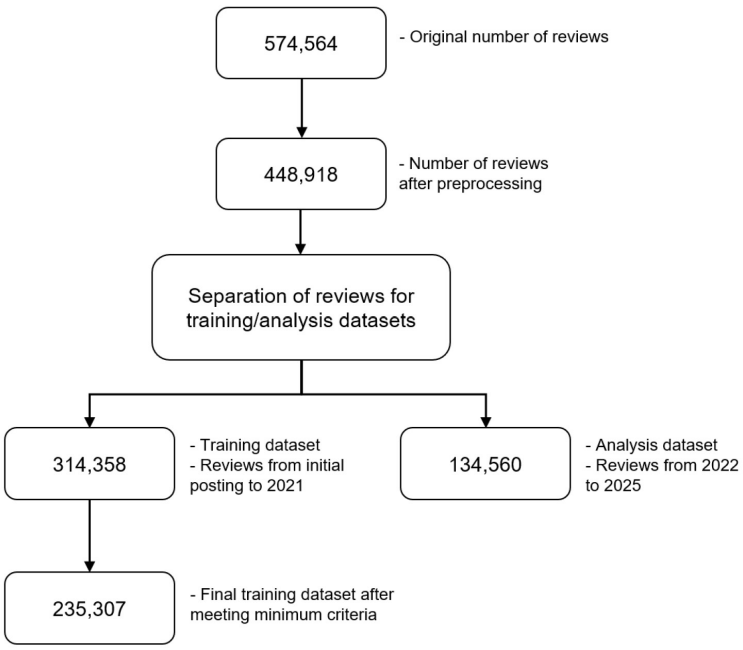

Figure 2. Overall Composition of Review Data for Training and Analysis

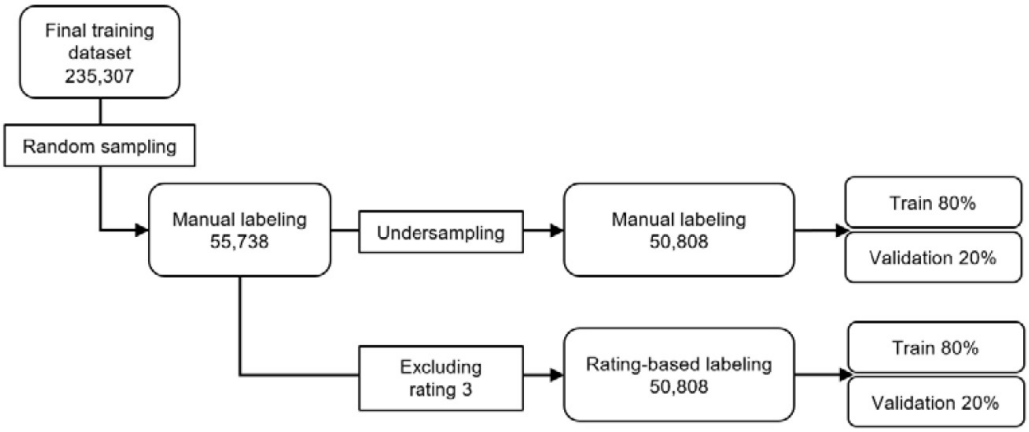

Figure 3. Construction Process of Training Datasets According to Each Labeling Method

동일한 리뷰 본문을 대상으로 평점 기반 라벨링을 수행하여 총 50,808건(긍정 25,404건, 부정 25,404건)의 학습용 데이터를 구축하였다. 이때 평점 1~2점은 부정, 4~5점은 긍정으로 간주하고, 중립 감정을 포함할 가능성이 높은 3점 리뷰는 분석에서 제외하였다. 이러한 평점 기준은 다수의 선행연구에서 리뷰의 평점을 기준으로 분류할 때 1~2점은 부정, 4~5점은 긍정으로 가정하는 방식에 근거한 것이다(Choi *et al.*, 2020). 결과적으로 중립 감정인 3점 리뷰가 분석에서 제외되면서 평점 기반 라벨링 학습용 데이터의 규모는 수작업 라벨링보다 적게 구축되었다. 따라서 두 방식 간 공정한 성능 비교를 위해 수작업 라벨링 학습용 데이터에 무작위 언더샘플링을 적용하여 평점 기반 라벨링과 동일한 수량으로 맞추었다. <Figure 3>은 각 라벨링 방식에 따른 학습용 데이터 구성 과정을 개략적으로 나타낸 것이다.

3.4 모델 학습

본 연구에서는 한국어 기반 감성분석 성능을 확보한 사전학습 모델을 기반으로, 학습 데이터 라벨링 방식에 따른 감성분석 성능 차이를 비교하고자 하였다. 이를 위해 N사 영화 리뷰 데이터셋인 NSMC을 활용하여 KoBERT 모델에 감성 개념을 학습시키는 1차 미세조정을 수행하였다. NSMC는 긍정과 부정의 이진 감정 라벨로 구성된 한국어 영화 리뷰 데이터셋으로 학습용 150,000건과 테스트용 50,000건으로 구성되어 있다. KoBERT는 방대한 코퍼스를 기반으로 한국어 문맥을 효과적으로 이해하고 처리할 수 있는데(Park *et al.*, 2024), NSMC 데이터셋을 학습용 데이터로 활용하였을 때 다른 LSTM 기반 모델보다 우수한 성능을 보였기 때문에(Hwang, 2021), 이를 1차 미세조정에 활용하였다. 1차 미세조정의 하이퍼파라미터는 <Table 3>에서 서술된 기존 연구(Lee *et al.*, 2023)의 조건을 참고하였다.

Table 3. Hyperparameter Settings used for the 1st Fine-tuning with NSMC

Parameters	1st fine-tuning using NSMC
max length	128
number of classes	2
batch size	32
optimizer	AdamW
epochs	10
learning rate	5e-5

2차 미세조정에서는 Batch size의 조절값에 따라 발생하는 노이즈의 양과 learning rate에 따라 달라지는 loss의 수렴이 감성분석의 정확도에 영향을 끼칠 수 있을 것이라고 판단한 기존 연구(Lee *et al.*, 2023)에 따라 Batch size 32와 64, learning rate 3e-5와 5e-5의 조합을 실험하였다. 각 조건별 결과는 Table 4에 제시하였으며, 그중 F1 score가 가장 높은 Batch size 64, learning rate 5e-5 조합을 최종적으로 선정하였다. 이후 해당 조건을 적용하여 수작업 라벨링 학습용 데이터와 평점 기반 라벨링 학습용 데이터로 각각 2차 미세조정을 수행하였다.

Table 4. Performance under Different Hyperparameter Settings for the 2nd Fine-tuning

Batch size 32			
Learning rate	Loss	Accuracy	F1 score
3e-5	0.401	0.825	0.826
5e-5	0.422	0.832	0.829
Batch size 64			
Learning rate	Loss	Accuracy	F1 score
3e-5	0.411	0.823	0.803
5e-5	0.421	0.832	**0.832**

두 가지 라벨링 방식으로 구축된 학습용 데이터는 8:2 비율로 훈련(train split)과 검증(validation split) 세트로 분할되었다. 이때

훈련 세트는 모델의 학습에 직접 사용되었으며 검증 세트는 학습 과정 중의 성능을 평가하고 과적합 여부를 확인하는 데에 사용되었다. 각 모델은 동일한 조건 하에서 무작위 Seed값을 1부터 30까지 변화시키며 반복 학습이 수행되었다. 성능 비교는 F1 score를 기준으로 평가하였는데, F1 score는 정밀도(Precision)와 재현율(Recall)의 조화 평균으로 계산되는 통계적 지표로, 정밀도와 재현율이 모두 높아야 큰 값을 나타내기 때문에 클래스 불균형 문제에서 중요 요소로 여겨진다(Yoo *et al.*, 2024). 이에 따라 각 라벨링 방식으로 학습된 KoBERT 모델의 평균 F1 score를 기준으로 성능을 비교하였다.

3.5 통계 분석

도출된 두 라벨링 방식 간 성능 차이의 통계적 검증을 위한 세 가지 분석을 수행하였다. 첫째, 무작위 Seed값을 1부터 30까지 변화시키며 반복 학습한 결과의 F1 score에 대해 대응표본 t-검정을 실시하여 두 라벨링 방식 간 평균 F1 score 차이가 통계적으로 유의한지 확인하였다. 둘째, 두 라벨링 방식 간의 도출된 예측 정확도(Accuracy)에 대해서도 대응표본 t-검정을 실시하였다. 두 모델은 동일한 조건에서 학습된 대응표본 관계이므로 대응표본 t-검정을 통해 평균 차이의 유의성을 평가하였다. 셋째, 동일한 정답 데이터셋을 기준으로 맥니마 검정을 실시하여 동일한 데이터셋에서 서로 다른 예측을 한 사례를 분석하여 두 모델 간 차이가 통계적으로 유의한지 평가하였다. 맥니마 검정에 사용된 데이터셋은 학습에 사용되지 않은 1,000건의 리뷰(긍정 500건, 부정 500건)를 연구자가 수작업으로 라벨링하여 구성하였다. 모든 통계 검정은 유의수준 $\alpha = 0.05$를 기준으로 실시하였다.

3.6 감성분석

모델 선정 후 2022년 리뷰 최초 작성 시점부터 2025년 4월 20일까지 총 134,560건의 리뷰로 구성된 분석용 데이터의 감성분석을 수행하였다. 각 리뷰는 긍정 또는 부정으로 분류되었으며, 각 금융앱별 긍정과 부정 리뷰 개수와 비율을 산출하였다. 전체 금융앱의 감성 분포 차이에 대한 통계적 검증을 위해 분산분석을 수행하였으며, 금융앱별 긍정 및 부정 리뷰 비율의 평균 차이가 유의미한지를 평가하였다. 이후 도출된 긍정과 부정 리뷰들을 통합하여 전체 금융앱을 포괄하는 긍정 리뷰 데이터셋과 부정 리뷰 데이터셋을 구축하였다. 두 개의 데이터셋은 토픽 모델링에 활용되었다.

3.7 토픽 모델링

감성분석을 통해 구축된 전체 긍정 리뷰 데이터셋과 전체 부정 리뷰 데이터셋을 대상으로 BERTopic 기반 토픽 모델링을 수행하였다. Table 5는 본 연구에서 사용된 BERTopic에서의 UMAP와 HDBSCAN의 주요 하이퍼파라미터 조건을 나타낸다. 해당 조건은 연구자가 주요 하이퍼파라미터의 값을 조정하며 비교한 결과, 주제 분리도와 해석 가능성이 가장 높은 조건을 최종적으로 채택한 것이다. Okt(Open Korean Text) 형태소 분석기를 사용해 명사 형태소를 추출하였으며, 본 연구의 목적을 고려하여 'UX', 'UI', 'GUI'와 같이 사용자 경험과 직결되는 주요 약어는 명사로 인식되도록 사전에 등록하였다. 또한 조사, 접속사, 감탄사, 의존명사 등 일반적인 불용어를 제거하였으며, '금융', '은행', '앱' 등 반복적으로 나타나지만 분석 목적에 부합되지 않는 단어들 역시 불용어로 간주하여 제외하였다. 한글 이외의 기호, 숫자, 영문자, 특수문자를 제거한 후, 의미가 유사하거나 혼용될 수 있는 표현들에 대해서는 단어 정규화(normalization)를 수행하였다. 위 조건을 통해 전체 긍정 및 부정 데이터셋의 토픽 모델링을 수행하여 주요 토픽과 키워드, 키워드별 랜덤 추출된 대표 리뷰 20개를 확보하였다. BERTopic의 특성상 HDBSCAN 클러스터링 과정에서 주제 군집이 명확하지 않은 항목은 Topic -1은 분류되기 때문에(Kim *et al.*, 2024b) 해석에서 제외하였고, 최종적으로 확보된 토픽과 키워드는 사용자 경험 관점과 허니콤 모델에 대응시켜 해석하였다.

Table 5. Hyperparameter Settings for UMAP and HDBSCAN in BER Topic

Algorithm	Parameter			
UMAP	n_neighbors	n_components	min_dist	metric
	30	5	0.0	cosine
HDBSCAN	min_cluster_size	min_samples	metric	-
	15	5	euclidean	-

4. 연구결과

4.1 라벨링

결과적으로 두 라벨링 방식 간 성능 차이는 모두 통계적으로 유의하지 않았다. F1 score 기준 대응표본 t-검정 결과, 수작

Table 6. Statistical Comparison of Model Performance by Labeling Method

Statistical test	Comparison metric	Manual labeling	Rating-based labeling	t / p-value
Paired t-test (F1 score)	Mean F1 score	0.868	0.873	-1.752 / 0.09
Paired t-test (Accuracy)	Mean accuracy	0.830	0.832	-0.198 / 0.84
McNemar's test	Number of mismatched cases	50	52	0.92

업 모델의 평균은 0.868, 평점 기반 모델의 평균은 0.873이었으며, p-value는 0.09로 유의하지 않았다($p \geq 0.05$). 예측 정확도에 대한 대응표본 t-검정 결과 역시 p-value 0.84로 통계적으로 유의하지 않았으며, 맥니마 검정 결과는 수작업 모델만 정답인 사례가 50건, 평점 기반 모델만 정답인 사례가 52건으로 나타났고, p-value 0.92로 차이가 없었다($p \geq 0.05$). 이에 대한 통계 분석 결과는 <Table 6>에 제시하였으며, 두 모델 간 성능 차이는 통계적으로 유의하지 않고 성능 측면에서 실질적인 차이가 존재하지 않는다는 결과가 도출되었다. 이에 본 연구에서는 감성분석에 활용되는 딥러닝 언어 모델의 학습 과정에서 수작업 라벨링이 요구하는 장시간의 작업 시간과 인적 자원 소모라는 한계점에 주목하였다(Park *et al.*, 2023). 평점 기반 라벨링 또한 리뷰 내용과 평점 간의 불일치 가능성이 존재하나(Kim and Song, 2016) 두 방식 간 유의한 성능 차이가 없었으며, 수작업 라벨링에 비해 상당한 시간과 비용을 절감할 수 있는 점에서 실용적일 것이라고 판단되었다. 따라서 본 연구에서는 평점 기반 라벨링 방식을 활용하여 학습한 모델을 최종 감성분석에 활용하였다.

4.2 감성분석

앞서 선정된 평점 기반 KoBERT 모델을 활용하여, 모바일 금융앱 25개에 대한 감성분석을 수행하였다. <Table 7>과 <Figure 4>는 학습 데이터 전처리 과정을 동일하게 거친 총 134,560건의 리뷰가 감성분석을 통해 긍정 76,028건(56%), 부정 58,532건(44%)으로 분류된 것을 정리한 것이다. 분산분석 수행 전 정규성 검정을 통해 정규성을 만족함을 확인하였으나, 등분산성 검정의 Levene 검정 결과에서는 등분산성이 충족되지 않았다($F = 215.57$, $p < 0.05$). 이에 Welch's ANOVA를 사용하였으며, 결과적으로 앱 간 긍정 리뷰 비율과 부정 리뷰 비율 모두 통계적으로 유의미한 평균차이가 있는 것으로 나타났다($F = 685.76$, $p < 0.001$). 도출된 긍정 리뷰 76,028건과 부정 리뷰 58,532건은 전체 긍정 리뷰 데이터셋, 전체 부정 리뷰 데이터셋으로 구축되어 토픽 모델링의 데이터셋으로 활용되었다.

Table 7. Sentiment Analysis Results for Each Financial App

No.	Financial sector	App name	Total review	Positive review	Negative review	Positive ratio	Negative ratio
1	Commercial banks	A	12,403	8,933	3,470	72%	28%
2		B	6,152	1,944	4,208	32%	68%
3		C	2,094	1,022	1,072	49%	51%
4		D	3,678	1,798	1,880	49%	51%
5		E	5,223	1,978	3,245	38%	62%
6		F	3,753	1,057	2,696	28%	72%
7	Regional banks	G	687	205	482	30%	70%
8		H	1,154	809	345	70%	30%
9		I	1,046	658	388	63%	37%
10		J	143	48	95	34%	66%
11		K	507	194	313	38%	62%
12	Internet-only banks	L	2,121	1,108	1,013	52%	48%
13		M	4,299	2,916	1,383	68%	32%
14	Fintech services	N	15,706	7,446	8,260	47%	53%
15		O	2,244	972	1,272	43%	57%
16		P	1,859	741	1,118	40%	60%
17	Credit card companies	Q	17,744	14,295	3,449	81%	19%
18		R	7,585	4,253	3,332	56%	44%
19		S	5,438	3,061	2,377	56%	44%
20		T	4,669	2,069	2,600	44%	56%
21		U	6,800	2,565	4,235	38%	62%
22		V	18,640	12,358	6,282	66%	34%
23		W	3,909	2,655	1,254	68%	32%
24		X	2,632	1,007	1,625	38%	62%
25		Y	4,074	1,936	2,138	48%	52%
Total			134,560	76,028	58,532		

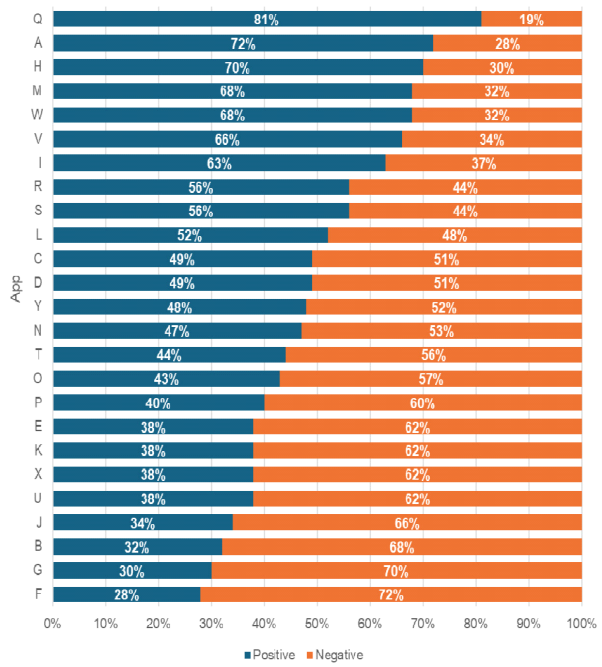

Figure 4. Positive and Negative Review Ratio by Financial App

4.3 토픽 모델링

전체 긍정 리뷰 데이터셋(76,028건)와 부정 리뷰 데이터셋(58,532건)를 대상으로 BERTopic을 적용하여 주요 토픽을 도출하였다. 추출된 키워드와 무작위로 선정된 키워드별 대표 리뷰 20건을 함께 검토하여 해석하였으며, 자의적 해석을 최소화하고자 허니콤 모델에 대응시켜 객관성을 확보하고자 하였다. <Table 8>에는 긍정 리뷰와 부정 리뷰별로 추출된 키워드와 대응되는 허니콤 모델의 범주를 정리하였다.

긍정 리뷰에서 나타난 상위 토픽은 다음과 같다. Topic 1은 '혜택', '최고', '정보', '보기', '업데이트' 등의 키워드와 "혜택 정보가 잘 정리되어 있다", "업데이트 후 보기 편해졌다"는 리뷰가 함께 나타나 서비스가 제공하는 가치성이 긍정적으로 평가되었음을 보여주었다. Topic 2는 '결제', '할인', '거래', '대출', '쿠폰' 등의 키워드와 "결제가 원활하다", "할인 쿠폰이 유용했다", "대출 거래가 간편하다"는 리뷰가 반복적으로 확인되었으며, 이는 효율적이고 손쉬운 이용과 관련된 사용성을 반영한다. Topic 3은 '메뉴', '찾기', '구성', '직관', '속도' 등의 키워드와 메뉴 탐색이 쉽다는 긍정적 반응이 다수 확인되어 정보 접근과 기능 탐색의 용이성을 보여주는 검색성이 강조되었으며, 동시에 서비스 기능을 원활히 접근할 수 있다는 점에서 접근성 역시 긍정적으로 평가되었다. Topic 4는 'UI', '인터페이스', 'UX', '사용자', '디자인' 등의 키워드와 함께 "디자인이 깔끔하다", "UI가 만족스럽다"는 평가가 나타났으며, 이는 심미적 만족과 감성적 요소에 해당하는 매력성과 연결된다. Topic 5는 '해외여행', '환전', '수수료', '결제' 등의 키워드와 "환전 수수료가 낮다", "해외 결제가 원활하다"는 반응이 반복되어 실제 생활에서의 편리함을 강조하는 유용성과 관련된다.

다음으로, 부정 리뷰에서 도출된 상위 토픽은 다음과 같다. Topic 1은 '비밀번호', '인증', '업데이트', '에러', '해외여행' 등의 키워드와 "비밀번호 및 인증이 계속 실패한다", "업데이트 이후 에러가 많다"는 불만이 반복되어 제공되는 서비스의 신뢰성이 저해되었음을 보여주었다. Topic 2는 '경험', '최악의', '인증', '사용자', '업데이트' 등의 키워드와 "최악의 사용자 경험이었다", "사용자를 전혀 배려하지 않는다"는 불만이 다수 확인되어 서비스가 제공해야 할 가치성이 충족되지 못했음을 시사하였다. Topic 3은 'UI', '화면', '메뉴', '인터페이스', '직관' 등의 키워드와 "UI가 복잡하다", "메뉴를 찾기 어렵다", "직관적이지 않다"는 의견이 반복되었으며, 이는 정보 접근과 기능 탐색 과정에서 검색성이 떨어지는 문제를 보여주었고, 더 나아가 사용자가 원하는 서비스에 원활히 접근하지 못한다는 점에서 접근성 측면의 한계도 드러났다. Topic 4는 '생체인식', '인식', '패턴', '통합', '인증' 등의 키워드와 "지문 인식이 자주 실패한다", "앱 통합 후 생체인식이 작동하지 않는다", "인증 방식

Table 8. Topic Modeling Results of Positive and Negative Reviews for Financial Apps with UX Honeycomb Model

Emotion	Topic	Honeycomb model	Keyword
Positive	1	Value	benefit, best, information, view, update
	2	Usability	payment, discount, transaction, loan, coupon
	3	Findability, Accessibility	menu, search, structure, intuitiveness, speed
	4	Desirability	UI, interface, intuitiveness, UX, user
	5	Usefulness	overseas travel, currency exchange, domestic, fee, payment
Negative	1	Credibility	password, signup, update, error, overseas travel
	2	Value	experience, worst, authentication, user, recognition
	3	Findability, Accessibility	UI, screen, menu, interface, user
	4	Credibility	biometric, pattern, integration, error, method
	5	Credibility	server, network, error, resolution, occurrence

변경 방법을 찾기 어렵다” 등의 불만이 확인되어 인증 과정에서의 신뢰성 저하로 해석할 수 있다. Topic 5는 ‘서버’, ‘네트워크’, ‘에러’, ‘해결’, ‘발생’ 등의 키워드와 “서버 오류가 잦다”, “통신 에러 발생”과 같은 불만이 다수 나타나, 서비스 인프라의 불안정성이 전반적인 신뢰성 문제로 연결됨을 보여주었다.

5. 토의

본 연구에서는 모바일 금융앱 리뷰를 기반으로 딥러닝 언어모델의 학습 데이터 구축 방식의 차이에 따른 모델의 감성분석 성능을 비교하고, 성능이 더 우수한 모델로 금융앱 리뷰의 감성분석을 수행하였다. 분석 결과를 토대로 토픽 모델링을 통해 긍정 및 부정 주요 UX 요소를 도출하여 피터 모빌의 허니콤 모델을 통해 긍정 리뷰에서의 강화 방안과 부정 리뷰에서의 개선 방안을 제시하였다.

학습 데이터 구축 방식에 따른 모델의 성능 차이를 비교하고자 수작업 라벨링 방식과 평점 기반 라벨링 방식을 모델에 반복 학습시킨 뒤 성능 차이를 비교하였다. 무작위 Seed값을 기반으로 반복 학습한 결과의 평균 F1 score과 평균 예측 정확도에 대한 대응표본 t-검정, 동일한 정답 데이터셋을 기반으로 한 맥니마 검정 결과, 두 모델 간의 통계적으로 유의한 차이는 나타나지 않아 학습 데이터 구축 방식에 따른 성능 차이는 실질적으로 유의하지 않음을 확인할 수 있었다. 그럼에도 본 연구에서 평점 기반 라벨링 모델을 최종 감성분석에 채택한 이유는 다음과 같다. 수작업 라벨링은 문맥을 고려한 정밀한 분류가 가능하다는 장점이 있으나, 연구자 간 해석 차이로 인해 감정 분류 기준의 일관성을 완전히 유지하기 어려웠다. 특히 한국어는 동일한 단어도 문맥에 따라 의미가 달라질 수 있어(Chae *et al.*, 2025), 키워드 해석의 주관성이 판단에 영향을 미쳤다. 본 연구에서도 라벨링 이견이 발생한 문장에 대해 별도의 재검토를 수행하였으나, 완전한 일관성을 달성하는 데는 한계가 있었다. 따라서 연구자의 주관성을 최소화할 수 있는 기준 마련(Lee *et al.*, 2021)과 교차 검증 등의 추가적인 보완 절차가 필요하다. 반면, 평점 기반 라벨링은 사용자가 직접 선택한 평점을 기준으로 하기 때문에 기준이 명확하고 자동화 가능성이 높다는 점에서 효율적이다. 일부 리뷰에서 평점과 본문 간 불일치가 존재할 수 있으나(Yun *et al.*, 2023) 분석 결과 평점 기반 모델은 수작업 기반 모델보다 평균 F1 score 수치가 더 높았다. 결과적으로 두 라벨링 방식은 통계적으로 유의한 성능 차이를 보이지 않았으나, 효율성, 확장성, 실용성, 정량적 성능을 종합적으로 고려할 때 평점 기반 라벨링 방식이 대규모 데이터를 다루는 감성분석 연구에 적합한 방식이라고 판단하였다.

평점 기반 모델을 활용한 감성분석 결과, 분석용 데이터 134,560건은 긍정 76,028건(56%), 부정 58,532건(44%)으로 분류되었다. 이는 모바일 금융앱에 대한 전반적인 사용자 경험이 비교적 긍정적으로 평가되고 있음을 보여주는 한편, 44%에 달하는 부정 리뷰는 여전히 다양한 불편과 문제점이 존재함을 시사한다. 전체 금융앱 간 감정 비율 차이에 대한 Welch’s ANOVA 결과, 긍정 및 부정 리뷰 비율 모두에서 통계적으로 유의미한 차이를 보였으며($p < 0.001$), 이는 동일한 금융 분야에 속하더라도 각 앱이 제공하는 서비스 품질에 따라 뚜렷한 편차가 존재함을 보여준다. 긍정 리뷰 비율이 가장 높은 앱은 81%를 기록한 반면, 가장 낮은 앱은 28%에 불과하였으며, 부정 리뷰 비율이 50% 이상인 앱도 전체 25개 중 15개에 달했다. 이러한 감정분석 결과의 정량적 편차는 각 앱의 사용자 경험 품질을 비교하고 진단하는 데 활용될 수 있는 기초 자료로서 의미가 있다. 특히 부정 감정이 집중된 앱을 식별함으로써, 사용자 불만이 지속적으로 축적되는 상황을 사전에 파악하고 대응할 수 있을 것이다. 감성분석 결과는 모바일 금융앱의 UX 문제를 사용자의 감성을 기반으로 파악할 수 있게 해주며, 사용자 경험의 질적 수준을 정량적으로 평가하는 실증적 접근에서의 활용 가능성을 보여준다.

도출된 감성분석 결과를 기반으로 토픽 모델링을 통해 긍정 및 부정 리뷰에서 나타나는 주요 UX 토픽들을 도출하였다. 도출한 토픽은 사용자 경험 전반에 대한 정서적 반응과 인식을 주제별로 구조화한 결과이며, 이를 통해 사용자가 실제로 만족하거나 불편함을 느끼는 구체적인 요인을 파악할 수 있었다. 또한 토픽들을 피터 모빌의 허니콤 모델의 유용성, 사용성, 탐색성, 매력성, 접근성, 신뢰성, 가치성 측면으로 해석하여 UX 요인을 체계적으로 분석할 수 있도록 하였다.

긍정 리뷰 분석 결과, 허니콤 모델의 가치성, 사용성, 검색성, 접근성, 매력성, 유용성 범주에서 사용자 만족이 두드러지게 나타났다. 가치성 측면에서는 제공되는 혜택과 정보 구조가 긍정적으로 평가되었으며, 이는 사용자 맞춤형 정보 제공과 체계적인 콘텐츠 배치의 중요성을 시사한 기존 연구 결과(Kim and Lee, 2013)와 일치한다. 또한 혜택에 대한 사용자의 관심도가 높기 때문에, 사용자의 관심사나 소비 패턴을 반영한 혜택 자동 추천 기능, 카드형 UI, 타임라인 기반 구성 등을 고려한다면 정보 탐색 효율 향상에 기여할 수 있을 것으로 기대된다(Lee and Kim, 2018). 사용성에서는 결제 및 금융 거래의 원활함이 강조되었고, 이러한 긍정적 요소의 강화를 위해 고객의 거래와 상호작용 데이터를 분석하여 안정적인 개인별 맞춤 서비스 제공을 고려해볼 수 있다(Park and Bae, 2024). 검색성과 접근성은 메뉴와 내비게이션 구조가 직관적일 때 사용자 만족을 높인다는 점을 반영하며, 정보 구조의 단순화와 탐색 경로의 명확성이 중요함을 의미한다. Zheng and Jung(2024)은 직관적인 내비게이션과 다양한 레이아웃 옵션을 위해 정기적으로 피드백을 수용하여 디자인 개선에 반영하는 것이 필요하다고 하였다. 하지만 일부 사용자의 숙련도나 사용 이력에 따라 경험 편차가 발생할 수 있으므로, 동적 메뉴 구성을 위한 사용성 실험과 초보자를 위한 튜토리얼 UX 제공(Lee and Ryu, 2024)이 유용할 수 있다. 매력성은 심미성이 높고 일관된 디자인이 사용자 경험을 향상시킨다는 점을 확인시켜 주었으며, 사용자 친화적인

인터페이스 제공을 통해 충성도 높은 고객을 확보하여 경쟁력을 키우는 것이 중요하다(Kong, 2025). 유용성은 해외 결제와 환전과 같은 부가 기능이 실질적인 편익을 제공할 때 사용자에게 긍정적으로 인식된다는 점을 보여주었다. 해외 결제와 환전의 편리함은 지갑을 대체할 수 있는 간편함을 의미하며, 서비스 이용 과정이 간편하고 편리하다고 생각한 사용자들은 지속적인 서비스 이용으로 이어지기 때문에(Jung *et al.*, 2019) 해외 결제 시스템의 확대를 고려해볼 수 있다.

부정 리뷰에서는 주로 신뢰성, 가치성, 검색성, 접근성 범주에서 문제점이 확인되었다. 신뢰성은 로그인 및 인증 오류, 생체인식 실패, 서버 불안정성과 같은 문제로 저하되었다. 서버 및 네트워크의 불안정성은 간편 결제와 같이 실시간성을 요구하는 서비스에서 치명적인 UX 결함으로 작용하며(Jung *et al.*, 2019), 시스템 품질 요인에서 보안성 및 연결성은 신뢰성에 유의한 영향을 미친다고 하였다(Koh *et al.*, 2014). 따라서 안정적이고 신뢰할 수 있는 인증 및 서버 인프라 관리가 필요함을 시사한다. 가치성은 최악의 사용자 경험이라는 서비스 가치 훼손으로 나타났다. 서비스 초기에 불만족스러운 경험을 한 사용자는 해당 앱을 재사용할 비율이 25% 미만이라는 기존 연구 결과에 따라(Hong and Pan, 2020), 실제적인 사용자 경험 단계와 가이드라인을 재검토하는 방안이 고려된다. 검색성, 접근성에서는 복잡한 메뉴와 직관적이지 않은 인터페이스가 불만을 초래했으며, 특히 감성분석 결과 부정적 리뷰 비율이 높았던 금융앱은 내비게이션 단순화와 정보 구조의 개선이 고려된다. 직관적이지 않은 정보 구조는 곧 복잡성과 직결되며, 이를 개선하기 위해 기능 검색의 강화, 시각적 강조, 사용자 수준 기반 메뉴 구성 등 사용자를 위한 직관적인 인터페이스 설계가 요구된다(Bae, 2021). 부정 리뷰에서의 전반적인 평가를 통해 각 금융앱은 개선 요소 보완을 위해 페르소나 기반 사용자 시나리오 재설계(Kim *et al.*, 2020) 및 전반적인 시스템 개편을 진단을 고려해볼 수 있다.

6. 결 론

본 연구는 모바일 금융앱 사용자 리뷰 데이터를 기반으로 수작업 라벨링과 평점 기반 라벨링 방식의 차이에서 감성분석 모델의 성능을 비교하고, 성능이 더 우수한 모델을 활용하여 감성분석을 수행하였다. 이후 토픽 모델링을 통해 긍정 및 부정 리뷰에 내재된 주요 UX 토픽을 도출하고, UX 강화 방안 및 개선 방안을 제시하였다.

라벨링 방식 비교 결과, 두 모델 간 성능 차이는 통계적으로 유의하지 않았다. 평점 라벨링 방식은 평점 기반으로 긍정과 부정을 판단하므로 자동화가 가능하며, 이에 따라 데이터 작업 시간이 수작업 라벨링 방식보다 크게 단축된다. 평점과 실제 리뷰 간의 간극으로 인한 일부 노이즈가 포함되더라도, 수동 라벨링 방식과의 성능 차이가 유의하지 않다는 점은 평점 라벨링 방식이 실무 환경에서도 효율적으로 작동할 수 있음을 시사한다. 이러한 결과는 대규모 학습 데이터 구축 시, 수동 라벨링 방식보다 빠르고 효율적으로 데이터를 구축할 수 있는 평점 기반 라벨링 방식의 유효성을 입증하며, 효율성과 정확도 간의 균형을 고려해 라벨링 방식을 선택할 수 있는 정량적 근거를 제공한다.

감성분석 결과, 전체 리뷰의 56%는 긍정, 44%는 부정으로 분류되었으며 앱 간 긍부정의 비율 차이가 통계적으로 유의하다는 것을 확인하였다. 이는 동일한 금융앱 분야에 속하더라도 각 금융앱에서 제공하는 사용자 경험은 명확한 편차가 존재함을 시사한다.

감성분석 이후 토픽 모델링을 통해 긍정 리뷰에서는 사용자 만족과 편의성에 기여하는 요소를, 부정 리뷰에서는 개선이 요구되는 문제 요소를 허니콤 모델을 기반으로 해석하였다. 이를 통해 제안된 모바일 금융앱 전반에 적용 가능한 UX 강화 방안 및 개선 방안을 <Table 9>에 정리하였다. 분석 범위는 시중은행, 카드사, 핀테크, 지역은행의 다양한 금융 분야를 포함

Table 9. UX Improvement Suggestions Based on Honeycomb Model Analysis of Positive and Negative Reviews

Emotion	Honeycomb model	UX suggestions
Positive	Value	Personalized information provision, Card-based UI, Timeline-based content organization
	Usability	Stable and fast service, Personalized service through analysis of transaction and interaction data
	Findability, Accessibility	Simplified information architecture, Clear navigation paths, Regular feedback integration, Dynamic menu configuration, Tutorial UX for beginners
	Desirability	Enhancing customer loyalty by providing a user-friendly interface
	Usefulness	Consider expansion of overseas payment system
Negative	Credibility	Stable and reliable authentication and infrastructure management
	Value	Re-examination of user experience stages and guidelines
	Findability, Accessibility	Simplified navigation, Improved information architecture, Enhanced search functions, User-level-based menu configuration
	Overall UX	Persona-based scenario design, User journey-based problem diagnosis

하였으며, 이를 통해 개별 앱에 국한된 것이 아닌 모바일 금융 앱 분야의 전반적인 개선안이라는 확장성을 확보하였다.

그럼에도 불구하고 본 연구는 다음과 같은 한계를 가진다. 첫째, 사용자 리뷰를 구글 플레이스토어 기반으로 수집했기 때문에, 아이폰 사용자나 타 플랫폼 이용자의 경험은 포함되지 못하였다. 둘째, 토픽 모델링에서 사용된 BERTopic 모델은 자동 클러스터링 기능에서 강점을 지니고 있으나(Kim *et al.*, 2024), 토픽 해석 과정에서 연구자의 주관이 개입될 수밖에 없다(Kwon *et al.*, 2024). 특히 유사 키워드 간 의미 중첩이나 대표 리뷰 해석의 다양성으로 인해, 동일 토픽에 대한 해석이 달라질 가능성이 존재한다. 마지막으로, 본 연구는 모바일 금융 앱 UX에 대하여 강화 방안 및 개선 방안을 제시하였으나, 도출된 방안의 실제 효과를 실증적으로 검증하는 절차는 수행되지 않았다. 향후 연구에서는 시선 추적 기반의 실험 등 실제 사용자 반응을 정량적으로 측정할 수 있는 설계를 도입하여, 제안된 UX 개선 요소가 사용자 만족도 및 사용성 향상에 미치는 영향을 검증하는 후속 연구가 필요하다.

참고문헌

Bae, D. E. (2021), *An Analysis on Determinants of Using Personal Financial Management Fin-Tech Application: Based on Text Mining for Banksalad Users*, Master's Thesis, Yonsei University, Korea.

BRI Korea (2025), https://brikorea.com/bbs/board.php?bo_table=rep_1&wr_id=4762 (retrieved April 25, 2025).

Chae, H. S., Lim, H. D., and Seo, Y. H. (2025), Word Sense Disambiguation Based on Word-Specific Classification Models, *Journal of the Korea Contents Association*, **25**(1), 15-20.

Cho, J. H. and Oh, H. Y. (2022), Training Techniques for Data Bias Problem on Deep Learning Text Summarization, *Journal of the Korea Institute of Information and Communication Engineering*, **26**(7), 949-955.

Choi, J. Y., Kim, H. A., and Kim, Y. B. (2020), The Impact of Online Review Volume, Rating, and Sentiment Score on Sales: Focusing on the Moderating Effect of Brand Reputation, *Journal of Channel and Retailing*, **25**(3), 1-21.

Jung, M. S. and Yeoun, M. H. (2022), User Experience Analysis of Financial MyData Service: Focusing on Comparison between Fintech and Traditional Banks, *Journal of Korean Society of Design Science*, **50**(4), 209-218.

Consumer Insight (2025), https://www.consumerinsight.co.kr/boardView?no=3708&id=pr18_list&PageNo=1&schFlag=0&viewFlag=1 (retrieved April 25, 2025).

Han, S. H. and Lim, J. H. (2014), UX Evaluation and Analysis of Smart Banking Service: Focused on Transfer Using the Method of Fingerprint Verification, *Journal of Integrated Design Research*, **13**(4), 33-48.

Hong, S. H. and Pan, Y. H. (2020), A Convergence Study on Initial User Experience of Mobile Banking Service, *Journal of Communication Design*, **73**, 631-646.

Hwang, S. H. (2021), Text Sentiment Analysis Based on Transformer Models Using an Emotional Dictionary, *Proceedings of the Korean Institute of Information Scientists and Engineers Conference*, Jeju, 2021.

Hyun, J. Y., Yoo, S. I., and Lee, S. Y. (2019), A Study on the Improvement of Recommendation System by Combining Ratings and Sentiment Analysis of Review Texts, *Journal of Digital Contents Society*, **20**(8), 1563-1573.

Irmawati, R., Chai, B., and Gunawan, D. (2022), Optimizing CNN Hyperparameters for Blastocyst Quality Assessment in Small Datasets, *IEEE Access*, **10**, 88621-88631.

Jan, S. L. and Shieh, G. (2014), Sample Size Determinations for Welch's Test in One-Way Heteroscedastic ANOVA, *British Journal of Mathematical and Statistical Psychology*, **67**(1), 72-93.

Jin, W. and Lee, J. W. (2022), Analyzing Game Streaming Application Reviews Using Text Mining Approach: Research to Strengthen Digital Competitiveness, *Journal of Digital Convergence*, **20**(4), 279-290.

Jo, H. J. and Chang, D. R. (2020), A Study on UI Design to Improve the Usability of Mobile Banking Applications for the Silver Generation, *Journal of Brand Design Association of Korea*, **18**(2), 137-149.

Jung, B. K. (2024). A study on analyzing the factors that influence composition methods using mobile device on music production workflow, *Asia-Pacific Journal of Convergent Research Interchange*, **10**(5), 107-118.

Jung, M. J., Lee, Y. L., Yoo, C. M., Kim, J. W., and Chung, J. E. (2019), An Exploratory Study on Consumers' Responses to Mobile Payment Service: Focused on Samsung Pay, *Journal of Digital Convergence*, **17**(1), 9-27.

Kang, S. W., Lim, H. G., and Son, H. S. (2024), Automatic Labeling Method for Overlapped Flatfish Object Detection Based on Image Processing Technique, *Smart Media Journal*, **13**(11), 59-70.

Kim, D. H., Park, H. J., Park, H. S., Lee, J. H., and Ko, S. S. (2024), Analysis of E-commerce App Service Positioning through Customer Reviews, *Journal of Big Data Service Studies*, **2**(1), 29-45.

Kim, E. M., Nam, S. J., Kim, T. Y., and Hong, T. (2024a), The Prediction of Review Helpfulness by Integrating Large Language Models and Deep Learning Based on KoBERT and KoGPT2, *Journal of Intelligence and Information Systems*, **30**(2), 195-209.

Kim, G. Y. and Han, S. M. (2022), User Review Analysis of English Learning Applications on Google Play Store Using Text-Mining, *Journal of Digital Contents Society*, **23**(10), 1901-1908.

Kim, H. K. and Hwang, W. Y. (2020), Proposal for Improving Data Processing Performance Using Python, *Journal of Korea Institute of Information, Electronics, and Communication Technology*, **13**(4), 306-311.

Kim, J. M. (2024), Trends in Crime Occurrence and Social Change: Focusing on Crime Data from 2012 to 2022, *Jungang Law Review*, **26**(4), 155-190.

Kim, J. M. and Chung, Y. J. (2024), Clustering Based Under-Sampling for Imbalanced Data Classification, *The Journal of Korean Institute of Information Technology*, **22**(5), 51-60.

Kim, N. Y., Kim, H. B., and Lee, S. G. (2025), Analyzing Foreign Tourists Experiences at Major Attractions in Seoul; Korea - Application of BERTopic Model Using Tripadvisor Reviews Data, *Journal of the Urban Design Institute of Korea Urban Design*, **26**(1), 51-70.

Kim, S. H., Song, Y., Lee, K. J., Putri, B., and Yun, M. H. (2020),

Design of Financial Customer Persona and Customer Journey Map Using Quantitative Day Reconstruction Method Data, *Proceedings of HCI Korea 2020*, Gangwon.

Kim, S. Y. and Kim, S. I. (2020), Evaluating the User Experience between Generations of Easy Payment Service: Focusing on KakaoPay and PAYCO, *Journal of Digital Convergence*, **18**(4), 453-459.

Kim, S. Y., Park, J. S., Lee, C. G., Yoo, J. W., Cho, Y. J., Yoo, J. Y., and Park, H. J. (2024b), Trends in Autonomous Vehicle Technology: Hierarchical Clustering Using BERTopic with a Focus on Patent Data Analysis, *The Journal of Intellectual Property*, **19**(4), 183-208.

Kim, Y. Y. and Song, M. (2016), A Study on Analyzing Sentiments on Movie Reviews by Multi-Level Sentiment Classifier, *Journal of Intelligence and Information Systems*, **22**(3), 71-89.

Koh, E. and Pan, Y. H. (2021), A Study on Financial Mobile Service UX/UI Audit Process, *Design Convergence Study*, **20**(6), 79-96.

Kwon, S. C., Kim, J. E., and Jang, B. C. (2024), Comparative Study of User Reactions in OTT Service Platforms Using Text Mining, *Journal of Internet Computing and Services*, **25**(3), 43-54.

Lee, C. H., Yun, Y. R., Bae, S. J., Eo, Y. D., Kim, C. J., Shin, S. H., Park, S. Y., and Han, Y. K. (2021), Analysis of Deep Learning Research Trends Applied to Remote Sensing through Paper Review of Korean Domestic Journals, *Journal of the Korean Society of Surveying, Geodesy, Photogrammetry and Cartography*, **39**(6), 437-456.

Lee, K. J. and Kim, S. I. (2018), Usability Evaluation for Simple Payment Service Based on Mobile Application: Focused on Shinhan and Samsung, *Journal of Digital Convergence*, **16**(9), 421-426.

Lee, M. A., Park, Y. J., Na, J. Y., and Sohn, C. B. (2023), Implementation of Review Sentiment Analysis Application Using KoBERT, KoGPT-2, and KoBART Optimized Hyperparameters, *Journal of Digital Contents Society*, **24**(11), 2831-2840.

Lee, S. J. (2024), An Analysis of Consumers' Responses toward E-Book Services Using LDA Topic Modeling and Sentiment Analysis Methods, *Journal of the Korea Contents Association*, **24**(10), 39-47.

Lee, T. K. and Ryu, H. K. (2024), A Study on the Improvement of Usability of Mobile Banking Credit Process Using MyData, *Journal of the HCI Society of Korea*, **19**(4), 5-12.

Lee, Y. L. and Choe, J. H. (2023), Proposed Guidelines for Improving UX Writing of Mobile Financial Platform: Focusing on the Cooperative Principle in Grice, *Asia-Pacific Journal of Convergent Research Interchange*, **9**(9), 675-683.

Li, W. Q. and Wu, Y. H. (2024), Research on Sentiment Analysis in Social Media App Reviews: Focusing on Instagram, *Journal of Emotional Science*, **27**(1), 69-80.

Majumdar, S., Kalamkar, S. D., Dudhgaonkar, S., Shelgikar, K. M., Ghaskadbi, S., and Goel, P. (2023), Evaluation of HbA1c from CGM Traces in an Indian Population, *Frontiers in Endocrinology*, **14**, Article 1264072.

Noh, H. J. and Son, C. H. (2021), Changes and Implications of the Financial Ecosystem Based on Open API, *KIRI Report (Focus)*, **521**, 8-13.

Park, K. W. (2022), Where do predictive biases in NLP models originate?, *Communications of the Korean Institute of Information Scientists and Engineers*, **40**(11), 30-35.

Park, M. Y., Kim, H. J., Lee, S. B., and Lee, J. W. (2023), Anomaly Detection and Root Cause Analysis of Ship Main Engines: Explainable Artificial Intelligence-Based Methodology Considering Internal Sensors and External Environmental Factors, *Journal of the Korean Institute of Industrial Engineers*, **49**(5), 379-394.

Park, U. H. and Bae, G. H. (2024), Super App Channel Strategy: A Case Analysis of Korean Commercial Banks, *Korean Management Consulting Review*, **24**(4), 219-229.

Park, W. C. (2025), A Study on an Early Diagnosis System for Children Language Developmental Disorders Using KoBERT, *Journal of the Korea Society of Computer and Information*, **30**(5), 51-58.

Rho, J. H. (2021), A Study on Preference of UX Cognitive Affordance Perspective for Mobile Simple Payment Users: Focusing on the Three Major Mobile Simple Payment Services in Korea, *The Treatise on the Plastic Media*, **24**(1), 202-210.

Shin, Y. S., Jeong, M. Y. and Lee, J. W. (2023), A Sampling Method for Resolving Semantic Redundancy and Bias in Autonomous Driving Image Datasets: Application to NIA AI Hub Dataset and Estimation of Annotation Cost Reduction, *KIISE Transactions on Computing Practices*, **29**(10), 451-466.

Son, J. I., Kim, Y. S., Noh, M. J., Pyo, G. J., Rahman, T., and Han, M. M. C. (2021), A Study on Classification of Mobile Application Reviews Using Deep Learning, *Smart Media Journal*, **10**(2), 76-83.

Yang, E. M. and Park, D. W. (2023), UI/UX Model of Multi-modal AI-based Mobile APP: Focusing on Senior-Friendly Services of Bank APPs, *Journal of the Korea Institute of Information and Communication Engineering*, **27**(9), 1037-1043.

Yoo, D. H., Kim, J. H., and Lee, H. C. (2024), Conditional Diffusion Model Based Data Augmentation Method for Efficient Classification in Class Imbalance Dataset, *The Journal of Korean Institute of Information Technology*, **22**(4), 79-90.

You, S. C., Choi, J. Y., and Sim, M. H. (2019). UX Design Evaluation and its Approach to Mobile Applications for Smart Appliances, *Smart Media Journal*, **8**(3), 70-79.

Yun, D. K., Park, K. T., and Choi, S. H. (2023), Development of a Tourist Satisfaction Quantitative Index for Building a Rating Prediction Model: Focusing on Jeju Island Tourist Spot Reviews, *Journal of Intelligence and Information Systems*, **29**(4), 185-205.

저자소개

박가율: 오산대학교 산업안전보건학과에서 2024년 학사 학위를 취득하고 아주대학교 산업공학과 석사과정에 재학 중이다. 연구 분야는 UX, 인간공학, 사용성 평가, 생체역학이다.

정명철 : Pennsylvania State University 산업공학과에서 2004년 박사학위를 취득하였다. 2005년부터 아주대학교 산업공학과 교수로 재직하고 있으며, 연구분야는 작업설계, 인간공학, 산업안전, UX/UI이다.

모승민: 한경국립대학교 안전공학과에서 2008년 학사 학위를 취득하고, 아주대학교 산업공학과에서 2010년 석사 학위, 2015년 박사 학위를 취득하였다. 현재 오산대학교 안전보건관리과 교수로 재직 중이다. 연구 분야는 산업안전, 인간공학, 생체역학, 인간-로봇 상호작용이다.