JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 30, No. 1: 1~12, Jan. 2026

토픽모델링 기반 고객 불만 분류 및 서비스 개선 방법 연구

원종수[1] · 배장원[2*]

Topic Modeling-based Customer Complaint Quality Classification and Service Improvement Strategies

Jong-Su Won[1] · Jang Won Bae[2*]

[1]Ph.D Student, Department of Industrial Management, Korea University of Technology and Education, Cheonan 31253 Korea
[2]*Associate Professor, Department of Industrial Management, Korea University of Technology and Education, Cheonan 31253 Korea

요 약

정보통신기업에서는 초고속 인터넷, IPTV, 유선전화 서비스의 품질 관리를 위해 고객 불만(Voice of Customer, VoC) 데이터 분석의 중요성이 지속적으로 증가하고 있다. 본 연구는 한국어 특화 언어모델인 KoBERT와 BERTopic을 결합하여 110,809건의 고객 불만 데이터를 분석하고, 이를 서비스 품질관리 체계에 연계하고자 하였다. 제안한 방법은 KoBERT를 활용해 상담 텍스트를 문장 단위 임베딩으로 변환한 후, BERTopic 기반 토픽모델링을 적용하여 고객 불만 유형을 자동으로 군집화하는 방식이다. 이를 통해 전체 데이터를 인터넷 품질, TV 품질, A/S 접수 후 정상, 유선전화 품질, 고객 자가 장비 및 환경 A/S의 5개 주요 서비스 품질 카테고리로 구조화하였다. 실험 결과, KoBERT 임베딩 모델은 정확도 0.9907, F1-score 0.9906을 기록하여 기존 한국어 언어모델 대비 우수한 성능을 보였다. 또한 BERTopic 모델링 결과 전체 문서의 99.7%가 토픽에 할당되어 높은 군집 품질이 확인되었다. 이러한 결과는 고객 불만 데이터를 기반으로 한 선제적 품질관리와 운영 효율성 향상을 위한 데이터 기반 의사결정 지원 가능성을 시사한다.

ABSTRACT

In the information and communications industry, analyzing customer complaints (Voice of Customer, VoC) is essential for managing the quality of high-speed Internet, IPTV, and fixed-line telephone services. This study integrates the Korean Bidirectional Encoder Representations from Transformers (KoBERT) model with Bidirectional Encoder Representations-based Topic Modeling (BERTopic) to analyze 110,809 customer complaint records and support service quality management. The proposed approach extracts sentence-level embeddings using KoBERT and applies BERTopic to automatically cluster complaint types. The dataset is structured into five service quality categories: Internet quality, TV quality, normal after A/S request, fixed-line telephone quality, and customer-owned equipment and environment A/S. Experimental results show that the KoBERT embedding model achieved an accuracy of 0.9907 and an F1-score of 0.9906, outperforming existing Korean language models. In addition, BERTopic assigned 99.7% of documents to topics, indicating high clustering quality. These results demonstrate that VoC-based topic modeling enables proactive quality management, data-driven decision-making, and improved operational efficiency in telecommunications services.

키워드 : 고객 불만, 서비스 품질관리, KoBERT, BERTopic, 토픽모델링

Keywords : Voice of Customer, Service Quality Management, KoBERT, BERTopic, Topic Modeling

Received 21 September 2025, **Revised** 3 November 2025, **Accepted** 17 December 2025

*** Corresponding Author** Jang Won Bae(E-mail:jangwon_bae@koreatech.ac.kr, Tel:+82-41-560-1443)
Associate Professor, Department of Industrial Management, Korea University of Technology and Education, Cheonan, 31253 Korea

Open Access http://doi.org/10.6109/jkiice.2026.30.1.1 pISSN:2234-4772

Ⅰ. 서 론

디지털 전환의 가속화와 네트워크 기반 서비스의 융합은 초고속 인터넷, IPTV, 유선전화(VoIP 포함) 등 정보통신 서비스의 품질관리 패러다임을 근본적으로 변화시켰다. 기존의 품질지표(고장률, 복구 시간, 신호 품질 등)만으로는 고객이 실제로 경험하는 서비스 품질과 만족도를 충분히 설명하기 어렵다. 특히 콜센터, 챗봇, 모바일 앱, 웹 포털 등 다중 접점에서 발생하는 대규모 비정형 텍스트 형태의 고객 불만(Voice of Customer, VoC)에 내재된 의미를 실시간에 가깝게 포착하고 품질 개선에 반영하는 역량이 핵심 경쟁 요소로 부상하였다.

이현정, 김향미, 이창섭(2020)은 통신서비스 분야에서 고객 불만 기반으로 한 실시간 모니터링이 고장 관리 및 선제적 품질 개선에 효과적이라고 보고하였다[1]. 또한 정보통신기업은 이전의 서비스 품질과 더불어 실시간 고객 서비스 분석을 통해 고객 민원 패턴의 사전 탐지가 가능하며 반복적인 고장의 구조적 개선으로 이어질 수 있음을 확인하였다[2]. 이러한 연구들은 서비스 품질관리가 더 이상 사후 서비스 준수 차원의 형식적 절차가 아니라, 데이터 분석과 자연어 처리 기반 선제적 서비스로의 전환이 필요함을 알 수 있다. 즉, 기존의 기술적 품질지표 중심 운영에서 벗어나, 고객의 소리를 체계적으로 분석하여 서비스 품질을 개선하려는 연구의 필요성이 더욱 커지고 있다.

특히 통신서비스는 장애 원인이 '망·단말·홈네트워크·플랫폼'으로 복합적으로 분산되어 있어, 정형 지표만으로는 고객 체감 품질 저하의 실제 원인을 규명하기 어렵다. 따라서 고객 불만 기반 비정형 텍스트를 정량적으로 구조화하여 품질관리 체계에 반영하는 과정은 서비스 장애의 근본 원인을 조기에 식별하고, 예방적 품질관리로 전환하는 데 필수적이다. 이러한 복합적 품질 요인 분석의 필요성은 선행 연구에서도 확인된다. 김명중, 김주일, 박선영(2017)은 IPTV 품질 인식에서 네트워크 품질(지연, 끊김, 버퍼링 등)이 고객의 부정적인 인식·이탈 의향에 유의한 영향을 미친다고 분석했고[3], 네트워크 품질·단말 품질·UI/UX 요인이 상호작용으로 작동함을 확인했다[4]. 이는 네트워크, 단말, 플랫폼 관점의 다차원적 품질 요인과 고객 발화(불만·요청·제안)의 의미를 함께 분석할 수 있는 통합적 프레임이 필요함을 시사한다.

고객 불만 데이터는 대량의 비정형 텍스트(상담 기록, 민원 내용, 채팅 로그 등)를 포함하며, 정형 메타데이터(상품·지역·회선 유형·고장 코드·처리 결과 등)와 결합될 때 고객 요구 처리 소요 시간 예측과 고객만족도 향상에 유용한 시사점을 제공할 수 있다[5]. 텍스트 마이닝, 감성분석을 통해 고객 불만의 핵심 속성을 정형화하고 예측 모형에 투입하면, 서비스 수준 협약서(Service Level Agreement, SLA) 준수율 제고나 선제적으로 조치해야 할 대상 식별 등 실질적 기여가 가능하다. 다만, 실제 기업 환경에서는 공백 데이터, 데이터 불균형, 도메인 용어(약자·오탈자·혼종어) 등 운영 데이터 특유의 문제가 존재하므로, 한글 도메인에 특화된 언어 모델과 강력한 클러스터링 모델의 결합이 요구된다.

최근 사전학습 언어 모델(Pre-trained Language Model, PLM)의 발전은 한국어 기반 비정형 텍스트 분석의 정확도를 크게 향상시켰다[6]. 특히 KoBERT는 한국어 형태·어절 특성을 반영한 SentencePiece 토크나이저를 기반으로 한국어 상담 데이터와 같이 구어체·혼종어가 포함된 텍스트를 효과적으로 임베딩할 수 있는 장점을 가진다[7]. BERTopic 또한 기존 확률모형 기반 토픽모델(Latent Dirichlet Allocation, LDA)의 제약을 보완하여, 다양한 도메인에서 주제 일관성과 해석 가능성을 높였다[8]. BERTopic은 문서 임베딩, 차원 축소(Uniform Manifold Approximation and Projection, UMAP), 밀도 기반 군집화(Hierarchical Density-Based Spatial Clustering of Applications with Noise, HDBSCAN), class-based Term Frequency-Inverse Document Frequency(c-TF-IDF)에 의해 주제 키워드를 도출한다. 이와 같은 토픽모델링은 정보통신기업의 고객 불만 데이터 분석에서 더욱 큰 의미를 갖는다. 서비스 불만, 장애 유형 분석, 장애 분류체계 고도화는 모두 텍스트 기반의 고객 상담 기록에 의존하는 특성이 있기 때문이다. 따라서 KoBERT-BERTopic 기반의 정량적 토픽 분석은 A/S 접수 단계에서의 오분류 감소, 중복 출동 방지, 문제 관리 프로세스의 선제적 대응 등 실제 서비스 품질 개선에 직결되는 도구로 활용될 수 있다. 이는 단순한 토픽 주제 도출뿐만 아니라 서비스 품질관리 체계 전반의 효율화를 가능하게 한다.

본 연구의 목적은 정보통신기업에서 발생하는 고객 불만 데이터를 KoBERT 임베딩과 BERTopic 모델링을 결합하여 분류하고, 이를 서비스 품질 개선을 통한 운영

비용 절감에 활용하는 데 있다. 총 110,809건의 고객 불만 상담 데이터를 분석하여 인터넷 품질, TV 품질, A/S 접수 후 정상, 유선전화 품질, 고객 자가 장비 및 환경 A/S 5개 핵심 카테고리로 분류하였다. 이를 통해 고객 불만 유형별 주요 패턴을 도출하고, 서비스 품질관리로 연계될 수 있는 기반을 마련하였다.

기존 연구가 주로 감성분석, 문서 분류 파악에 머문 것과 달리 본 연구는 정보통신기업의 A/S 카테고리와 토픽을 직접 매핑하여 현업 운영에 즉시 활용 가능한 모델을 제시한다는 점에서 차별성이 있다. 또한, 최신 연구들이 BERTopic의 일반적 적용 가능성을 검증하는 수준에 머물렀다면, 본 연구는 정보통신서비스의 복합적 구조를 고려하여 '토픽'과 '서비스' 품질 카테고리를 연결한 '품질관리 지향 토픽모델링'을 제시한다는 점에서 기존 연구와 차별성을 지닌다. 실험 결과, KoBERT 임베딩은 정확도 0.9907, F1-Score 0.9906, AUC-ROC 0.9806의 수준을 보였으며, BERTopic 모델링에서는 인터넷 품질 불만이 49.6%(54,912건), TV 품질 불만이 24.2%(26,822건)로 나타나 전체 불만의 73.8%가 인터넷·IPTV 영역에 집중되어 있음을 확인하였다. 또한 모델의 토픽 일관성을 나타내는 Coherence Score는 0.4737로 평가되어, 도출된 토픽 구조가 실제 서비스 품질관리에 활용 가능한 수준임을 검증하였다. 이는 고객 불만 데이터의 체계적 분류를 통해 서비스 운영상 주요 관리 영역을 식별하고, 품질 개선 및 비용 절감 전략에 직접 연계할 수 있는 기반을 제공한다는 점에서 의의가 있다.

Ⅱ. 관련 연구

2.1 토픽모델링과 품질관리 연구

품질관리는 전통적으로 고장률, 복구 시간, 불량률 등 정형 지표를 중심으로 운영되었다. 최근 비정형 데이터(고객 불만 등)의 실시간 데이터 분석을 통해 예방적 품질관리로 전환하려는 사례가 증가되고 있다[9].

비정형 텍스트를 품질관리에 통합하려는 시도는 크게 두 가지 경로로 진행된다. 첫째, 텍스트로부터 반복적·잠재적 품질 이슈를 탐지하고, 분류하여 우선순위를 정한 뒤 표준운영절차에 반영하는 것이다. 둘째, 고객 불만 텍스트 분석 결과를 정량적 품질지표와 결합하여 개선 조치의 효과성을 평가하고 검증하는 프로세스를 설계하는 것이다. 최근 연구에서는 BERTopic이 문서 임베딩과 차원 축소, 밀도 기반 군집화, c-TF-IDF 기반 키워드 도출을 결합함으로써 기존 LDA 계열보다 더 응집력 있고 해석 가능한 토픽을 도출한다는 점을 확인하였다. 특히 차원 축소와 밀도 기반 군집화의 조합은 다양한 비정형 데이터에서 유의미한 문서 군집을 효과적으로 분리해, 실무 적용 가능성을 크게 높이는 것으로 보고되었다[10,11]. 소프트웨어 개발 분야에서는 BERTopic을 활용하여 결함 리포트의 토픽을 도출하고 이를 다출력 분류기와 결합하여 결함 유형과 원인을 예측하는 시스템 구축을 통해, 결함에 기반한 예방적 품질관리를 구현하였다[12]. 서비스 품질 분야에서는 항공사 고객 리뷰 데이터를 분석한 연구에서 고객이 인식하는 가치 요소를 토픽 단위로 분류하고, 이를 고객 만족도와 연계·분석하여 항공 서비스 운용자들이 고객 경험 기반의 차별화된 품질 전략을 수립할 수 있었다[13].

2.2 KoBERT 적용 연구

한국어 비정형 텍스트는 조사, 어미, 띄어쓰기, 혼종어 등의 특성으로 인해 영어권 모델을 그대로 적용하면 의미 손실과 토픽 품질 저하가 발생할 우려가 있다. 이를 보완하기 위해 한국어에 특화된 KoBERT, KR-BERT 등을 문장과 문서 임베딩 단계에 적용한 뒤 BERTopic을 구동하는 연구가 증가하고 있다[14].

한국어 특화 모델을 도메인 텍스트(민원, 상담 내용, SNS) 전처리 및 임베딩에 활용하면 문맥 보존이 개선되어 군집의 응집도와 주제어의 유의미성이 증가한다[15]. 응용 연구로는 KoBERT를 적용 기업 관련 뉴스 기사와 감성분석 정확도를 테스트하여 KoBERT 모델의 분류 정확도가 85.5%로 다국어 버전인 다른 모델들보다 높은 정확도를 보였다[16].

따라서 KoBERT 모델 활용은 한국어 고객 불만 기반 품질관리에서 토픽의 의미적 정밀도와 주제 응집도를 높이고, 실시간 고객 불만 유형의 자동화된 탐지·분류가 가능한 분석 프레임워크로 적합함을 알 수 있다.

2.3 BERTopic 빅데이터 분석 연구

최근 빅데이터 분석 기법으로서 BERTopic의 활용이 다양한 산업에서 증가하고 있다. 헬스케어 분야에서는 병원 환자 피드백 데이터를 BERTopic으로 분석하여 핵

심 환자 불만 토픽들을 도출하고, 이를 서비스 품질 개선 및 의료 서비스 대응 전략 수립에 활용하였다[17]. 항공 안전 분야에서도 BERTopic을 활용해 안전 보고서로부터 주요 사고 유형과 패턴을 도출함으로써, 전통적인 잠재 의미 분석(Probabilistic Latent Semantic Analysis, PLSA)보다 토픽 응집도와 해석 가능성이 우수함(Coherence Score 0.41)을 입증했다[18]. 금융 소비자 보호(Financial Protection Bureau) 데이터를 대상으로 한 연구에서는 잠재 디리클레 할당(Latent Dirichlet Allocation, LDA) 및 잠재 의미 분석(Latent Semantic Analysis, LSA)보다 BERTopic이 더 의미 있고 다양한 토픽을 생성했으며, 특히 도메인 특화 임베딩을 활용했을 때 더욱 뛰어난 성과(Coherence Score 0.33)를 확인했다[19]. 또한, 아이디어 관리 및 브레인스토밍 세션에서 BERTopic을 도입해 아이디어 병합과 토픽 도출 과정을 자동화하면서, 의미론적으로 응집된 주제 생성을 가능하게 한 의사결정 지원 시스템을 제안했다[20]. 표 1은 이러한 내용을 종합 정리하였다.

이처럼, BERTopic은 단순한 텍스트 분석을 넘어, 다양한 분야의 품질 및 서비스 관리에 시사점과 운영 개선을 제공할 수 있는 효과적인 분석 수단으로 사용할 수 있다.

Table. 1 Previous Studies on BERTopic Applications for Quality Improvement

Author (year)	Method	Key Findings
Sajjadul Islam et al. (2025)	BERTopic	Derived key patient complaint topics for healthcare service improvement.
Nanyonga et al. (2025)	BERTopic vs. PLSA	BERTopic showed higher coherence and interpretability than PLSA.
Sangaraju et al. (2022)	BERTopic + FinBERT	Generated more meaningful and diverse financial topics than LDA/LSA.
Cheddak et al. (2024)	BERTopic	Consolidated ideas into cohesive clusters for innovation-driven quality management.

Ⅲ. 연구 방법

3.1 연구 절차

본 연구는 정보통신기업에서 발생하는 대규모 고객 불만 비정형 텍스트 데이터를 KoBERT 임베딩과 BERTopic 모델로 분석하여, 도출된 결과를 서비스 품질관리에 체계적으로 활용하고자 한다. 그림 1은 고객 불만 데이터가 상담센터에서 처리되어 A/S 요청으로 이어지는 업무 흐름을 보여주며, 이후 그림 2의 연구 절차에 따라 4단계 분석이 수행된다.

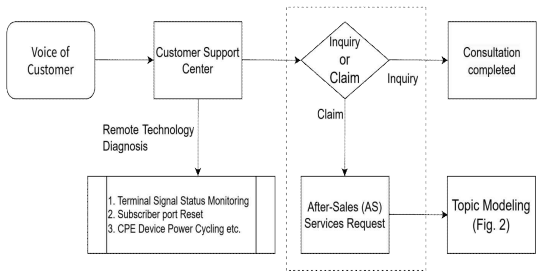

Fig. 1 Workflow of VoC and A/S Processing

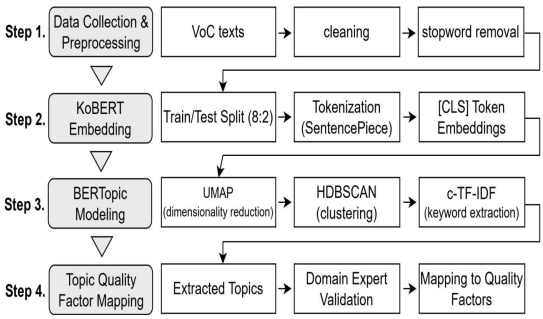

Fig. 2 Research Procedure Diagram

1단계 : 데이터 수집 및 전처리로 정보통신기업의 초고속 인터넷, IPTV, 유선전화 서비스 영역에서 수집된 대규모 고객 불만 데이터를 연구에 활용한다. 상담 및 불만 접수 내용 등 텍스트 데이터를 확보한 후, 불필요한 특수문자 제거, 띄어쓰기, 도메인 불용어(stopwords) 제거 등의 전처리 과정을 수행한다.

2단계 : KoBERT 임베딩 기반 의미 표현을 위해 한국어 특화 사전학습 언어 모델인 KoBERT를 활용하여 고객 불만 문장의 의미를 벡터화한다. 모델 성능 향상을 위해 고객 불만 상담 데이터셋을 기반으로 추가적인 학

습 및 미세조정(fine-tuning) 과정을 수행한다. 이를 통해 상담 대화와 민원 문장의 맥락적 의미를 정교하게 표현할 수 있으며, 한·영 혼합 표현(ssid 재연결, 구형 Wi-Fi 등) 및 도메인 특화 용어(giga, snmp, Wi-Fi, VOD 등)의 의미도 반영한다.

3단계 : BERTopic 기반 토픽 구조화 및 대표 키워드 도출로 KoBERT 임베딩 결과를 입력으로 차원 축소 - 밀도 기반 군집화 - c-TF-IDF로 구성된 BERTopic 모델을 적용한다. 모델의 토픽 분류 성능을 높이기 위해 밀도 기반 군집화 등 주요 단계별 하이퍼파라미터에 대한 최적화 탐색을 병행한다. 이 과정을 통해 유사 문서 군집을 형성하고, 각 군집에서 주제어를 대표하는 키워드를 도출하여 토픽 구조를 생성한다.

4단계 : 토픽별 A/S 품질관리 카테고리인 인터넷 품질, TV 품질, A/S 접수 후 정상, 유선전화 품질, 고객 자가 장비 및 환경 A/S로 매핑한다. 도출된 토픽과 각 품질관리 요소 간의 연관성과 해석 가능성을 높이기 위해 도메인 전문가 검증을 한다. 이 검증 과정을 통해 현업 적용 가능성을 높이고, 보완이 필요한 토픽에 대해서는 추가적인 분석 및 수정을 진행한다. 이를 통해 고객 불만 기반 토픽 구조가 실제 현업 품질관리 체계와 연계되도록 한다.

3.2 연구 방법

3.2.1 데이터셋과 전처리

본 연구에서는 국내 메이저 정보통신기업의 고객 불만 접수 시스템에서 직접 추출한 원본 데이터를 활용하였다. 해당 데이터는 상담원이 고장 상담을 실시간으로 입력·저장한 기록에 기반하며, 총 110,809건으로 구성되어 있다. 개인정보 데이터는 비식별 처리되어 있고, A/S 처리 결과가 함께 기재되어 있다. 본 데이터셋은 모두 88개 항목의 정형 필드로 구조화되어 관리되고 있으나, 분석의 핵심 대상은 상담원이 직접 작성한 비정형 텍스트 필드 '상담 내용'으로, 고객 불만의 실제 맥락 파악과 자연어 처리 기법 적용에 적합하다.

전처리 과정으로 상담 내용 필드에 결측값을 제거하고, 특수문자 및 불필요한 기호 제거와 상담 데이터의 특성상 불용어가 다수 포함되어 있어 분석의 왜곡 방지를 위해 일반 불용어(예: '이', '그', '저', '그리고' 등)뿐만 아니라, 상담 현장에서 반복적으로 나타나는 특수 용어(예: '역량', '센터', '진단 툴 확인' 등)를 포함한 확장형 불용어 처리로 불필요한 잡음을 최소화하였다.

전처리된 텍스트는 UTF-8 인코딩 형태의 문장 단위 문자열로 변환되어 KoBERT 모델의 SentencePiece 토크나이저를 통해 토큰화되며, 각 문장은 768차원 임베딩 벡터로 변환한다. 이후 해당 벡터들은 N x 768 행렬 형태로 구성되어 차원 축소, 밀도 기반 군집화, c-TF-IDF 기반 키워드 도출 단계의 입력 데이터로 활용된다.

3.2.2 KoBERT 임베딩

한국어 문장의 의미적 특징을 벡터로 변환하기 위해 사전학습된 KoBERT 모델을 활용한 임베딩을 수행하였다. 전체 고객 상담 텍스트를 학습(train) 데이터셋과 평가(test) 데이터셋 8:2비율로 분할하였다. 이후 GluonNLP 라이브러리의 SentencePiece 토크나이저를 활용하여 텍스트를 토큰화하였다.

KoBERT 분류 모델은 기본적인 BERT 구조 위에 선형 분류기(Linear Classifier)를 추가하여 구현되었으며, 교차 엔트로피 손실(Cross-Entropy Loss)을 지표로 사용하였다. 최적화에는 AdamW 옵티마이저와 코사인 웜업(cosine warm-up) 스케줄러를 적용하였고, 성능 평가는 정확도, 정밀도, 재현율, F1-Score 및 AUC-ROC 지표를 기준으로 진행되었다.

모델 학습에 사용된 주요 하이퍼파라미터는 표 2와 같이 최대 토큰 길이 256, 배치 크기 64, 학습률 5e-5, Epoch 25로 데이터 특성과 모델의 학습 안정성을 고려하여 실험적으로 도출하였다. 최대 토큰 길이 256는 상담 내용 문맥 손실을 최소화하기 위해 적용하였고, 배치 크기 64는 학습 안정성과 연산 효율 간의 균형을 고려해 탐색한 결과이며, 학습률 5e-5는 BERT 계열 모델에서 권장되는 범위 내에서 F1-Score가 가장 안정적으로 나타난 구간이었다. Epoch 수는 초기 학습 구간(1~15회차)에서 지속적인 성능 향상이 나타나고, 이후 완만히 수렴하는 패턴을 보였기 때문에 25 Epoch로 설정하여 학습 추이 분석 및 성능 포화 시점을 확인하였다. 이러한 탐색적 검증 결과를 통해 본 설정이 과적합을 방지하면서도 안정적인 분류 성능을 유지함을 확인하였다.

Table. 2 KoBERT Hyperparameter Settings

Hyperparameter	Value
Maximum Token Length	256
Batch Size	64
Learning Rate	5e-5
Epochs	25
Evaluation Metrics	Loss, Accuracy, Precision, Recall, F1 Score

학습이 완료된 후에는 KoBERT의 [CLS] 토큰 벡터를 추출하여 문장 단위의 의미 임베딩으로 활용하였다. 이렇게 도출된 임베딩은 이후 토픽모델링(BERTopic) 단계의 입력으로 사용되며, 상담 내용 텍스트의 의미적 유사성을 효과적으로 반영하는 벡터 표현으로 기능한다.

3.2.3 BERTopic 기반 토픽모델링

고객 불만 상담 텍스트 데이터의 잠재적인 주제 구조를 도출하기 위해 BERTopic을 적용하였다. BERTopic은 임베딩 벡터를 차원 축소한 뒤 클러스터링을 수행하고, 각 군집의 대표 키워드를 도출하는 방식으로 토픽모델링을 구현한다. 효과적인 토픽모델링을 위해 밀도 기반 군집화의 최적 하이퍼파라미터 조합을 탐색하는 과정을 거쳤다. 우선, KoBERT를 통해 생성된 임베딩을 UMAP(Uniform Manifold Approximation and Projection) 알고리즘을 활용하여 저차원 공간으로 축소하였다. 이를 통해 고차원 임베딩의 밀집 구조를 보존하면서도 효율적인 군집화를 가능하게 하였다. 차원 축소된 벡터는 이후 HDBSCAN(Hierarchical Density-Based Spatial Clustering of Applications with Noise) 알고리즘을 적용하여 밀도 기반 클러스터링을 수행하였다. HDBSCAN은 데이터 밀도 기준으로 클러스터를 식별하고, 잡음(outlier)으로 분류되는 문서를 별도로 처리하여 보다 안정적인 주제 구조를 도출할 수 있도록 한다.

마지막으로 각 클러스터 내에서 자주 등장하는 핵심 용어를 도출하기 위해 c-TF-IDF(class-based Term Frequency-Inverse Document Frequency) 방식을 적용하여 특정 토픽을 대표하는 차별적인 키워드를 효과적으로 도출한다.

본 연구에 적용한 BERTopic의 주요 하이퍼파라미터는 표 3과 같다. UMAP은 BERTopic의 기본 하이퍼파라미터(n_neighbors=15, n_components=5, min_dist=0.0, metric=cosine)를 사용하였으며, HDBSCAN은 min_cluster_size=350, min_samples=12, metric=euclidean, cluster_selection_method=eom으로 설정하였다. 또한 c-TF-IDF는 ngram_range=(1,1), min_df=3, max_df=0.7, max_features=10000을 적용하였으며, MMR(Maximal Marginal Relevance, diversity=0.8)을 설정하였다. 이러한 하이퍼파라미터 조합은 여러 실험을 통해 도출된 결과로, 각 단계별 토픽 품질지표(Coherence, Topic diversity, 할당률)를 비교·검증하여 결정되었다. 특히 UMAP의 n-neighbors=15는 문서 간 노이즈를 완화하기 위한 최적값으로 확인되었고, HDBSCAN의 min_cluster_size=350은 군집 응집도와 토픽 수의 균형을 동시에 확보하기 위한 설정이다. c-TF-IDF의 ngram_range=(1,1)과 min_df=3은 짧은 상담 문장 내 핵심 단어 중심의 해석력을 높이기 위한 조합이며, MMR diversity=0.8은 중복 키워드 억제를 통해 토픽별 대표어의 다양성을 확보하기 위한 기준으로 채택되었다. 결과적으로 본 설정은 Coherence Score 0.4737과 Topic Diversity 0.8400의 성능을 보였다.

Table. 3 BERTopic Hyperparameter Settings

Module	Hyperparameter=value
UMAP	n_neighbors=15,n_components=5, min_dist=0.0, metric=cosine
HDBSCAN	min_cluster_size=350, min_samples=12, metric=euclidean, cluster_selection_method=eom
c-TF-IDF	ngram_range=(1,1), min_df=3, max_df=0.7, max_features=10000, MMR(diversity=0.8)

이와 같은 과정을 통해 도출된 토픽은 문서 집합을 주제별 체계적으로 구조화하며, 각 토픽에 대한 대표 키워드와 관련 문서 분포가 함께 도출되어 고객 불만의 주요 패턴을 파악하는 데 활용된다.

3.2.4 토픽과 품질 요인 매핑

BERTopic으로 도출된 결과를 정보통신기업의 주요 품질 요인과 대응시키는 과정을 수행하였다. 각 토픽의 대표 키워드를 품질 항목으로 매핑하였다. 매핑 신뢰성을 높이기 위해 도메인 전문가를 통해 토픽의 핵심 키워드를 검토하고 해당 요인과의 연관성을 검증하였다. 본 연구에서 설정한 5개 품질 카테고리는 정보통신기업에서 운영되는 서비스 품질관리 체계 및 서비스 불만 분류

기준을 기반으로 선정하였다. 이 기준은 서비스 유형별 구조(예: 인터넷, TV, 유선전화)와 고객 상담에서 발생하는 불만 내용(예: 네트워크 접속 오류, VOD 지연, 끊김 등)을 반영하여 구성되었다. 본 연구에서는 도출된 토픽과 품질 요인을 아래와 같이 대응시켰다.

• Topic 0 → 인터넷 품질 : 네트워크 접속 오류, ssid 재연결, Wi-Fi 공유기 불량 등 인터넷 서비스 품질과 직결되는 항목.

• Topic 1 → TV 품질 : IPTV 시청 중 입력신호 오류, VOD 지연, 리모컨 동작 불능 등 IPTV 품질 이슈와 관련된 항목.

• Topic 2 → 고객 자가 장비 및 환경 A/S : 리모컨, 리모컨 버튼 불량 등 고객이 직접 사용하는 소모품 장비 고장으로 기인한 품질 문제를 포괄하는 항목.

• Topic 3 → A/S 접수 후 정상 : 고객이 불편을 신고했으나 이후 상담원과 1차 조치 또는 확인 시 '정상'으로 판정되는 사례를 반영하는 항목.

• Topic 4 → 유선전화 품질 : 팩스 송수신 불가, 라인 단절 등 연결 불량, 끊김, 등 유선전화 서비스의 품질 문제와 관련된 항목.

이와 같은 매핑 과정을 통해, 단순한 텍스트 기반 토픽모델링 결과를 실제 품질관리 체계의 관점에서 구조화할 수 있었으며, 품질 요인별로 고객 불만의 주요 패턴을 해석하고 관리 지표로 활용할 수 있는 기반을 마련했다.

Ⅳ. 실험 결과

본 연구에서는 정보통신기업의 고객 불만 데이터를 KoBERT 임베딩과 BERTopic 모델링을 통해 분석하고, 도출된 토픽을 기반으로 서비스 품질관리에 적용을 검증하였다. 4.1절에서는 고객 불만 상담 데이터를 대상으로 한 KoBERT 임베딩 성능을 평가하였으며, 모델의 학습 안정성을 확인했다. 4.2절에서는 BERTopic 모델링을 수행하여, 4.2.1절에서 토픽 분포 및 키워드 분석을 통해 주요 서비스 품질 문제를 식별하였고, 4.2.2절에서는 토픽 관계성 분석을 통해 토픽 간 상호 연관성과 그룹화 구조를 도출하였다. 이러한 일련의 분석 과정을 통해 본 연구는 고객 불만 데이터에서 의미 있는 토픽을 체계적으로 분류하고, 이를 서비스 품질 개선과 연계할 수 있는 실증적 근거를 제시한다.

4.1 KoBERT 임베딩 성능

정보통신기업의 고객 불만 상담 데이터 110,809건을 전처리하여 KoBERT 기반 임베딩을 진행하였다. 전처리 결과 평균 문서 길이는 23.4개 단어(최소 1, 최대 277개)로, 실제 상담 현장의 비정형 텍스트의 다양성을 반영하였다. 빈 문서 비율은 0%로 데이터 전처리 과정에서 결측 및 누락이 발생하지 않아 분석 품질이 확보되었음을 의미한다.

본 연구에서는 KoBERT 모델의 성능을 다양하게 검증하기 위해 정확도, F1-Score, AUC-ROC 등 다중 분류 모델에서 일반적으로 활용되는 평가 지표를 적용하였다. 또한 학습 안정성과 과적합 여부를 확인하고자 Epoch별 Training Loss와 정확도, 정밀도, 재현율, F1-Score의 변화를 함께 분석하였다. 이러한 평가 절차를 통해 모델의 학습 품질을 충분히 확인한 후, 고객 불만 데이터를 768차원의 문장 벡터로 변환하는 데 활용하였다. KoBERT 임베딩은 텍스트의 의미적 맥락과 도메인 특화 용어까지 효과적으로 반영하며, 성능 결과는 표 4와 같다. Epoch 25에서 Cross-Entropy Loss 0.0559를 기록했으며, 정확도 0.9907, F1-Score 0.9906, AUC-ROC는 0.9806의 성능을 나타냈다. 특히, KoBERT를 활용한 기존 한국어 감성 분류 연구에서 보고된 정확도 85.7%[16]와 비교할 때, 본 연구의 모델이 통신 도메인의 고객 불만 분류 작업에 적합하게 학습되었음을 입증한다.

Table. 4 KoBERT Model Performance Evaluation Results

result	KoBERT
Cross-Entropy Loss	0.0559
Accuracy	0.9907
F1-Score	0.9906
AUC-ROC	0.9806

아울러, 정확도와 F1-Score가 거의 동일한 수준(0.9907, 0.9906)을 기록했다는 점은 고객 불만을 균형적으로 학습했음을 시사한다. 또한 AUC-ROC 0.9806은 모델의 전반적인 구분 능력이 매우 우수함을 보여주어, 실제 서비스 운영 환경에서도 다양한 불만 유형을 높은 신뢰도로 판별할 수 있음을 확인할 수 있었다.

그림 3은 25 Epoch에 걸쳐 KoBERT 모델의 학습 성능을 보여준다. 학습 손실(Training Loss)이 지속적으로 감소하는 경향을 나타내며, 이는 모델이 시간이 지남에 따라 예측 오류를 최소화할 수 있음을 의미한다. 또한, 정확도, 정밀도, 재현율, F1-Score의 성능 지표들이 꾸준히 상승하여 1.0에 근접함을 보여준다. 15번째 Epoch 이후로는 성능 지표의 개선 폭이 0.001 미만으로 수렴하여, 안정적인 성능 수준에 도달했음을 알 수 있다.

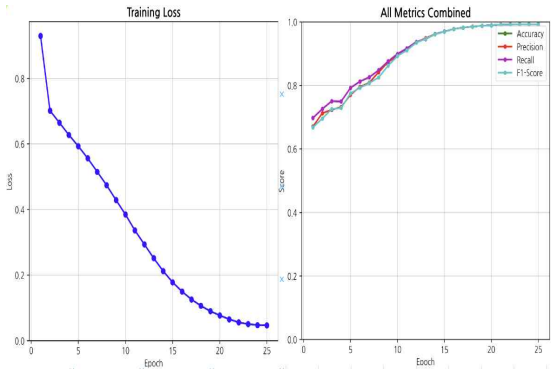

Fig. 3 KoBERT Training Metrics by Epoch

이러한 분석 결과는 KoBERT 임베딩 기반 토픽모델링 절차가 신뢰성 있는 품질 요인을 도출하기 위한 기반으로 적합함을 보여준다.

4.2 BERTopic 모델링

4.2.1 토픽 분포 및 키워드 분석

BERTopic 모델링을 통해 그림 4와 같이 총 5개의 주요 토픽이 도출되었으며, 전체 문서의 99.7%가 토픽으로 할당되는 군집 품질을 보였다. 각 토픽별 문서 분포는 Topic 0(54,912), Topic 1(26,822), Topic 2(20,954), Topic 3(7,365), Topic 4(397), 아웃라이어(-1)는 359건이다. 또한 모델의 토픽 일관성을 나타내는 Coherence Score는 0.4737로 평가되었다. 토픽모델링의 Coherence Score는 일반적으로 0.3 이상일 때 의미 있는 토픽이 도출되었다고 평가되며, 특히 BERTopic을 활용한 선행연구들 항공 안전 분야 0.41[18], 소비자 금융 보호 분야 0.32와 비교할 때[19], 본 연구의 0.4737은 더 높은 토픽 일관성과 해석 가능성을 확보했음을 의미한다.

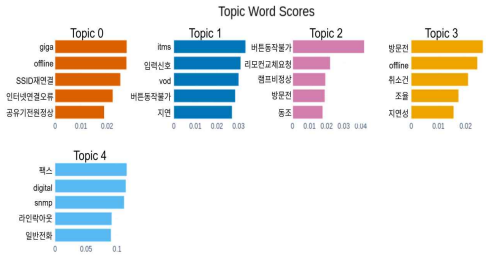

Fig. 4 Barchart of Topic Modeling Word Scores

Topic 0 : giga(0.028), offline(0.028), ssid 재연결(0.025), 인터넷 연결 오류(0.022), 공유기 전원 정상(0.020), 구형 wifi(0.018), 싱크 정상(0.018), 무료 전환(0.018), 기술(0.018), 싱크 비정상(0.017)과 같은 키워드가 높은 비중을 차지하고 '인터넷 품질'에 해당하는 토픽이다. 이들 키워드는 주로 초고속 인터넷 품질 저하 및 접속 불가 상황을 반영하며, 특히 공유기 전원이나 Wi-Fi 신호 재연결과 같은 고객 단말 및 네트워크 접속 문제를 포함하고 있다. 즉, 인터넷 품질 저하는 단순 회선 문제뿐만 아니라 공유기, SSID(Service Set Identifier) 인증, Legacy 장비 등 복합 요인에 의해 발생되고 있음을 보여준다.

Topic 1 : itms(0.033), 입력신호(0.032), vod(0.030), 버튼 동작 불가(0.029), 지연(0.027), 정상(0.024), 편성표(0.021), 업데이트(0.019), 영상(0.019), 블랙(0.018)과 같은 키워드가 높은 비중을 차지하고 'TV 품질'에 해당하는 토픽이다. 이는 IPTV 서비스 품질과 직결된 신호 오류, 영상 재생 지연, VOD 콘텐츠 접속 불량, 리모컨 입력 고장 등을 반영한다. 즉, Topic 1은 IPTV 시청 환경에서 발생하는 콘텐츠 품질 저하와 단말 동작 불능 문제를 포괄적으로 보여주며, 방송 서비스 품질관리의 중요성을 시사한다.

Topic 2 : 버튼 동작 불가(0.042), 리모컨 교체요청(0.022), 램프 비정상(0.019), 방문전(0.019), 동조(0.018), 작동불(0.016), 교체해도(0.016), 버튼 동작불(0.015), 정상이나(0.014), 이어폰(0.014)과 같은 키워드가 높은 비중을 차지하고 '고객 자가 장비 및 환경 A/S'에 해당하는 토픽이다. 이는 주로 고객이 사용하는 리모컨, 이어폰과 같은 고객 소유의 장비 불량에 해당한다. 특히 반복적인 리모컨 교체 요청과 작동 불가 증상은 고

객 환경에서 발생하는 단순 장치 불량임을 보여주며, 이는 현장 출동보다는 고객 안내 및 소모품 교체 정책을 통해 사전적으로 해결할 수 있는 영역으로 해석된다.

Topic 3 : 방문전(0.026), offline(0.024), 취소건(0.021), 조율(0.017), 지연성(0.016), 인터넷 연결 오류(0.015), 가능성(0.015), 입력신호(0.014), 지연(0.013), 자급(0.013)과 같은 키워드가 높은 비중을 차지하고 'A/S 접수 후 정상'에 해당하는 토픽이다. 이는 고객 불만 접수 이후 콜센터 상담사와 고객의 초기 대응을 통해 서비스 정상 복구 또는 자동 복구로 인한 예약 취소, 단순 조율, 재확인 과정에서 정상으로 판정되는 사례임을 보여준다. 이러한 토픽은 고객이 문제를 인지했으나 상담 후 확인 과정에서 정상 상태로 판정되거나 경미한 원인으로 해결되는 경우가 많음을 의미한다. 따라서 불필요한 A/S 출동을 줄이기 위한 상담원의 1차 진단 정확도 향상과 고객 자가 조치 안내 강화가 요구된다.

Topic 4: 팩스(0.116), digital(0.114), snmp(0.112), 라인 락아웃(0.091), 일반전화(0.090), 라인 확인(0.088), 송수신 먹통(0.085), 수화기 방치(0.079), 방문 전(0.073), 집 전화(0.061)과 같은 키워드가 높은 비중을 차지하고 '유선전화 품질'에 해당한다. 주로 회선 단절에 의한 팩스 송수신 불가와 같은 전화 서비스 품질 문제를 반영한다. 특히 digital과 snmp(Simple Network Management Protocol)는 디지털 교환기 및 네트워크 관리 시스템을 통한 회선 상태 모니터링과 관련된 기술적 요인을 반영하고 있다. 이러한 키워드 분포는 유선전화 서비스 품질이 단순한 음성 통화뿐만 아니라 팩스 통신, 회선 상태 관리, 디지털 신호 처리 등 다층적인 기술 요소들과 밀접하게 연관되어 있음을 보여준다.

본 연구의 BERTopic 기반 토픽 분석 결과는 단순히 기존 서비스 품질 문제의 유형을 재확인하는 데 그치지 않고, 각 토픽이 실제 고객 불만 처리 및 서비스 개선 정책 설계에 실질적으로 연결될 수 있음을 실증적으로 보여준다. 특히, 인터넷 품질 저하가 단순 회선 문제가 아니라 공유기, SSID 인증, 노후 Wi-Fi 장비 등 복합적인 원인임을 구체적으로 드러내었으며, TV 품질 영역에서는 입력신호 문제와 VOD 콘텐츠 접속 불량 관련 고객 불만의 중요성을 인식하였다. 또한 고객 자가 장비 및 환경 A/S 영역에 대한 토픽 도출을 통해 현장 출동이 아닌 고객 안내와 소모품 교체 정책 개선을 통해 대응 가능함을 확인했다. 마지막으로, 유선전화 품질 문제에서는 교환기 및 네트워크 관리 요소가 주요 원인임을 알 수 있었다. 종합하면, 본 연구의 결과는 서비스 품질 이슈별 원인과 대응 전략을 정교하게 제시함으로써 실무 적용 가능성을 높였다는 점에서 의의가 있다.

4.2.2 토픽 관계성 분석

토픽 관계성을 나타내는 거리 맵(Distance Map)은 BERTopic 모델 내부에서 계산된 토픽 벡터 간의 코사인 유사도(Cosine Similarity)를 기반으로 한다. 각 토픽을 대표하는 토픽 벡터가 c-TF-IDF를 통해 생성되며, 이 토픽 벡터들이 얼마나 유사한지를 코사인 유사도로 측정한다. 이후, 고차원 유사도 행렬을 차원 축소를 사용하여 2차원 평면에 시각화한 것이 그림 5와 같다. 두 토픽이 가까울수록 토픽을 구성하는 키워드 및 문서들이 의미적으로 높은 유사성을 가지며, 이는 두 토픽이 서비스 품질 관점에서 상호 연관되어 있음을 의미한다. 또한, 표 5와 같이 전체 토픽이 두 개 주요 그룹으로 명확히 구분됨을 확인할 수 있었다.

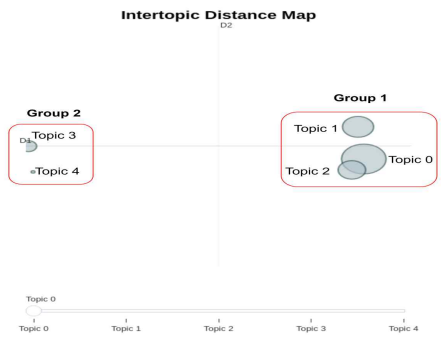

Fig. 5 Distance Map of Topics

Table. 5 Group of Topic due to Relationship

Group	Topics
1	Topic 0 : Internet Quality Topic 1 : TV Quality Topic 2 : Customer Device & Environment AS
2	Topic 3 : A/S Normalized after Request Topic 4 : Fixed-line Telephone Quality

첫 번째 그룹(Group 1)은 'Topic 0(인터넷 품질)', 'Topic 1(TV 품질)', 'Topic 2(고객 자가 장비 및 환경 A/S)'로 구성된다. 이 그룹은 고객이 체감하는 서비스

품질 저하와 직접적으로 연관되어 있다. 인터넷 접속 불가, IPTV 영상 지연, 리모컨 불량 등은 고객 체감 품질 저하로 즉시 인식되며, 상담원의 초기 진단이나 리모컨과 같은 소모품 교체 등 1차 대응 절차와 밀접히 연결된다. 특히 인터넷과 IPTV 서비스는 Wi-Fi 공유기, 셋탑박스, 리모컨 등 고객 단말 및 부속 장비와 밀접히 연관되어 있어, 동일한 그룹 내에서 연관성을 보인다. 따라서 이러한 그룹은 고객 체감 품질 및 단말/소모품 환경 중심 문제로 요약할 수 있다.

두 번째 그룹(Group 2)은 'Topic 3(A/S 접수 후 정상)', 'Topic 4(유선전화 품질)'로 구성된다. 이 그룹은 상대적으로 서비스 운영 절차와 네트워크 망 관리 특성이 밀접하게 연관된다. A/S 접수 후 정상으로 판정되는 사례는 기술적 장애보다는 작업으로 인한 끊김, 일시적 오류, 자동 복구, 고객 서비스 안내 부족에서 비롯되는 경우다. 이는 상담 프로세스와 고객 소통 체계의 효율성과 관련된다. 반면 유선전화 품질은 팩스 송수신 불가, 회선 단절, SNMP 기반 관리 이슈와 같은 기술적 품질 문제를 반영한다. 따라서 이 그룹은 서비스 운영 절차 및 네트워크 관리 문제로 요약할 수 있다.

종합하면, 그룹 내부에서는 토픽 간 연관성이 높게 나타나 동일한 관리 체계로 접근할 수 있지만, 그룹 간에는 고객 환경 중심 문제와 네트워크 관리 문제라는 차이가 존재한다. 이는 정보통신기업의 서비스 품질관리가 단일 프로세스로 해결되기보다, ① 고객 단말 및 환경 관리 체계와 ② 서비스 운영 및 네트워크 관리 체계라는 이원적 구조로 병행 운용될 필요성을 시사한다.

Ⅴ. 결 론

본 연구는 정보통신기업의 초고속 인터넷, IPTV, 유선전화 서비스 영역에서 발생하는 대규모 고객 불만 데이터를 기반으로, 한국어 특화 언어 모델 KoBERT와 BERTopic 기반 토픽모델링을 적용하여 서비스 품질관리를 위한 시사점을 도출하고자 하였다. 기존의 서비스 품질관리는 주로 정형 데이터(고장률, 복구 시간, 품질 지표 등)에 의존해 왔으나, 실제 고객이 경험하는 서비스 품질 문제는 상담 및 불만 접수 내용 등 비정형 데이터 속에 잠재되어 있어 이를 효과적으로 분석할 필요가 있었다. 따라서 고객 불만 기반의 비정형 텍스트를 구조화하여 품질관리 체계와 연계하는 분석 접근은 서비스 품질을 정교하게 이해하는 데 매우 중요한 수단이 된다.

본 연구의 제안 방법은 다음과 같다. 먼저, 고객 불만 데이터를 KoBERT를 이용해 문장 단위 임베딩을 수행하였다. 이후, 생성된 임베딩을 BERTopic 모델을 통해 차원 축소, 군집화를 거쳐 토픽을 자동으로 도출하였다. 마지막으로, 도출된 토픽을 A/S 카테고리와 매핑하여 품질관리 관점에서 분석하였다.

연구 결과, 총 5개의 주요 토픽이 도출되었으며, 전체 문서의 99.7%가 토픽으로 분류되었다. 토픽별로는 인터넷 품질, TV 품질, A/S 접수 후 정상, 유선전화 품질, 고객 자가 장비 및 환경 A/S로 분류되어 서비스의 핵심 불만 요인을 반영하는 주제가 확인되었다. 각 토픽의 대표 키워드 분석을 통해 인터넷 접속 불가, IPTV 입력신호, 리모컨 불량, 상담 과정 중 정상 확인, 유선전화 회선 단절 등 구체적 문제가 나타났으며, 이는 고객 불만의 주요 패턴을 체계적으로 이해할 수 있었다. 토픽별 문서 분포에서도 인터넷 품질 49.6%(54,912건)와 TV 품질 24.2%(26,822건) 관련 불만이 전체의 약 73.8%를 차지하여, 인터넷·IPTV 중심의 품질 개선 우선순위가 실증적으로 확인되었다.

토픽 간 관계성을 분석한 결과, 두 개의 그룹이 형성되었다. 첫 번째 그룹은 인터넷 품질, TV 품질, 고객 자가 장비 및 환경 A/S로 구성되어 고객 체감 품질 문제 및 고객 내 단말, 소모품 환경 요인이 반영되었다. 두 번째 그룹은 A/S 접수 후 정상, 유선전화 품질로 구성되어 서비스 운영 절차와 네트워크 관리 특성을 보였다. 이러한 결과는 정보통신기업의 서비스 품질관리가 단일 프로세스가 아니라, ① 고객 단말 및 환경 관리 체계와 ② 서비스 운영 및 네트워크 관리 체계라는 이원적 구조로 관리되어야 함을 실증적으로 알 수 있다. 즉, 고객 체감 품질 대응과 망 안정성 확보라는 두 가지 전략이 동시에 요구된다.

연구의 학술적 의의는 다음과 같다. 첫째, 감정 분석이나 일반적 토픽 분류 연구와 달리 실제 A/S 카테고리와 직접적으로 연계된 토픽 구조를 제시함으로써 품질관리 체계와의 정합성을 확보했다는 점이다. 둘째, 한국어 특화 언어 모델 KoBERT와 BERTopic을 결합하여 한국어 고객 불만 데이터 특유의 혼종어·약어·구어체를 정밀하게 반영했다는 점에서 기존 연구 대비 차별성을 가진다. 셋째, 토픽 간 관계성 분석을 통해 불만 유형 간

의 구조적 연관성 및 그룹화 패턴을 확인하여 품질관리 정책 설계에 활용 가능한 객관적 근거를 제시하였다.

연구의 실무적 의의는 다음과 같다. 첫째, 상담 단계에서 '정상'으로 귀결되는 사례를 분석함으로써 불필요한 A/S 출동 감소와 운영 비용 절감을 도모할 수 있다. 둘째, 인터넷 및 IPTV 불만 비중이 높다는 결과는 예방적 유지보수, 장비 교체 정책, 네트워크 안정화 전략의 우선순위 설정에 직접적 근거가 된다. 셋째, 리모컨 등 소모품 불량은 출동 없이 정책적 교체 및 고객 안내 강화로 효율적 대응이 가능함을 확인하였다. 넷째, 유선전화 품질 문제는 교환기 및 회선 관리 강화의 필요성을 제기하며 SLA 준수율 제고에 기여할 수 있다.

본 연구의 한계점으로는 특정 정보통신기업의 데이터에 한정되어 있으며, 데이터 불균형으로 일부 토픽의 성능이 낮게 나타난 점이 있다. 향후 연구에서는 다수 기업의 데이터를 활용한 교차 검증, 최신 언어 모델과의 비교 실험, A/S 결과 데이터와의 통합 분석을 통해 모델의 일반화 가능성과 실무 활용 범위를 확장할 필요가 있다.

REFERENCES

[1] H. Rhee, H. Kim, and C. S. Rhee, "Transition of Service Paradigm from Service Recovery to Proactive Service," *The Journal of the Korea Contents Association*, vol. 20, no. 4, pp. 396-405, Apr. 2020. DOI: 10.5392/JKCA.2020.20.04.396.

[2] M. Gewaly, M. S. Saeed, and M. A. Ammar, "Machine Learning Approach for Complaint Prediction in the Telecom Industry," in *Proceedings of the 2024 Intelligent Methods, Systems, and Applications (IMSA)*, Giza: EG, pp. 12-17, 2024. DOI: 10.1109/IMSA61967.2024.10652677.

[3] M. J. Kim, J. Kim, and S. Y. Park, "Understanding IPTV churning behaviors: focus on users in South Korea," *Asia Pacific Journal of Innovation and Entrepreneurship*, vol. 11, no. 2, pp. 190-213, Aug. 2017. DOI: 10.1108/APJIE-08-2017-026.

[4] Y. S. Kang, G. H. Kim, and J. G. Kim, "The Empirical Analysis on IPTV Customer Dissatisfaction: in Terms of System Quality," *Journal of The Korean Data Analysis Society*, vol. 13, no. 4, pp. 2153-2163, Aug. 2011.

[5] J. Kim and O. Kwon, "A Method of Predicting Service Time Based on Voice of Customer Data," *Journal of Information Technology Services*, vol. 15, no. 1, pp. 197-210, Mar. 2016. DOI: 10.9716/KITS.2016.15.1.197.

[6] M. Kim, Y. Kim, Y. Lim, and E. N. Hun "Advanced Subword Segmentation and Interdependent Regularization Mechanisms for Korean Language Understanding," in *Proceedings of the 2019 Third World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)*, London: GB, pp. 221-227, 2019. DOI: 10.1109/WorldS4.2019.8903977.

[7] SK Telecom Open Source. KoBERT [Internet]. Available: http://sktelecom.github.io/project/kobert/.

[8] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," *arXiv preprint arXiv: 2203.05794*, Mar. 2022. DOI: 10.48550/arXiv.2203.05794.

[9] C. U. Pyon, J. Y. Woo, and S. C. Park, "Intelligent service quality management system based on analysis and forecast of VOC," *Expert Systems with Applications*, vol. 37, no. 2, pp. 1056-1064, Mar. 2010. DOI: 10.1016/j.eswa.2009.06.066.

[10] L. M. Innes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv: 1802.03426*, Feb. 2018. DOI: 10.48550/arXiv.1802.03426.

[11] L. M. Innes, J. Healy, and S. Astels, "hdbscan: Hierarchical density based clustering," *Journal of Open Source Software*, vol. 2, no. 11, pp. 1-2, Mar. 2017. DOI: 10.21105/ oss.00205.

[12] D. P. Gottumukkala, P. V. G. D. Reddy, and S. K. Rao, "Topic modeling-based prediction of software defects and root cause using BERTopic, and multioutput classifier," *Scientific Reports*, vol. 15, no. 25428, Jul. 2025. DOI: 10.1038/s41598-025-11458-0.

[13] E. J. Jeong, B. H. Lee, Q. L. Li, and J. K. Kim, "The Effect of Perceived Customer Value on Customer Satisfaction with Airline Services Using the BERTopic Model," *Knowledge Management Research*, vol. 24, no. 3, pp. 95-125, 2023. DOI: 10.15813/kmr.2023.24.3.006.

[14] S. G. Lee, H. S. Jang, Y. M. Baik, S. Z. Park, and H. P. Shin, "Kr-bert: A small-scale korean-specific language model," *arXiv preprint arXiv: 2008.03979*, Aug. 2020. DOI: 10.48550/arXiv.2008.03979.

[15] K. Yang, "Transformer-based Korean pretrained language models: A survey on three years of progress," *arXiv preprint arXiv: 2112.03014*, Nov. 2021. DOI: 10.48550/arXiv.112. 03014.

[16] J. W. Hyeon, J. I. Lee, and H. K. Cho, "Sentiment analysis of news on corporation using KoBERT," *Korean Accounting Review*, vol. 47, no. 4, pp. 33-54, Aug. 2022. DOI: 10.24056/KAR.2022.08.002.

[17] K. M. S. Islam, R. T. Karri, and S. Vegesna, J. Wu, P. Madir aju, "Contextual Embedding-based Clustering to Identify To pics for Healthcare Service Improvement," in *Proceedings of the 49th Annual Computers, Software, and Applications Con*

ference (COMPSAC), Toronto: CA, pp. 794-799, 2025. DOI: 10.1109/COMPSAC65507.2025.00106.

[18] A. Nanyonga, J. Keith, T. Ugur, and W. Graham, "Is BERTopic Better than PLSA for Extracting Key Topics in Aviation Safety Reports?," *arXiv preprint arXiv: 2506.06328*, May 2025. DOI: 10.48550/arXiv.2506.06328.

[19] S. Vasudeva Raju, B. Kumar Bolla, D. K. Nayak ,and J. Kh, "Topic modelling on consumer financial protection bureau data: An approach using BERT based embeddings," in *Proceedings of the 7th International conference for Convergence in Technology (I2CT)*, Mumbai: IN, pp. 1-6, 2022. DOI: 10.1109/I2CT54291.2022.9824873.

[20] A. Cheddak, T. A. Baha, Y. E. Saady, M. E. Hajji, and M. Baslam, "BERTopic for enhanced idea management and topic generation in Brainstorming Sessions," *Information*, vol. 15, no. 6, pp. 365, Jun. 2024. DOI: 10.3390/ info15060365.

원종수(Jong-Su Won)

2015년 한국기술교육대학교 IT융합과학경영학과 과학경영석사
2022년~현재 한국기술교육대학교 산업경영학과 박사과정
※관심분야 : 자연어 처리, ICT, 기술경영

배장원(Jang Won Bae)

2007년 고려대학교 전기전자전파공학 공학사
2009년 한국과학기술원 전자공학과 공학석사
2015년 한국과학기술원 산업및시스템공학과 공학박사
2020년~현재 한국기술교육대학교 산업경영학과 부교수
※관심분야 : 모델링, 시뮬레이션, 빅데이터, AI