중간보고서

텍스트와 이미지 기반 패션 리뷰 분석 자동화 모델 개발

정보대학원 비즈니스 빅데이터 분석 트랙

황성아(2023522119), 박세연(2023522114)

https://github.com/SungaHwang/Text-Image_DLProject

1. 연구 필요성 및 배경

1.1 연구 필요성

패션 산업은 쇼핑몰 시장의 중요한 분야로, 소비자의 참여와 피드백에 의존한다. 특히 코로나19 이후 쇼핑몰 시장의 경쟁력이 성장함과 동시에 치열해지고 있으며, 이에 따라 패션 쇼핑몰 회사는 사용자의 피드백을 빠르게 반영하여 서비스를 개선하는 것이 중요하다.

패션 플랫폼 시장에서 큰 점유율을 가진 M사에서는 하루에도 수많은 사용자 리뷰가 업로드 되며, 이는 제품 개선과 소비자의 의사결정에 중요하다. 사용자 리뷰는 제품에 대한 직접적인 사용자 경험과 반응을 담고 있어, 제품의 장단점을 파악하고 의사결정의 방향성을 찾을 수 있다. 패션 분야에서는 사용자 리뷰가 색상, 사이즈, 매칭 가능한 스타일 등 실제 착용 시의 다양한 요소에 대한 평가를 포함하므로 이를 분석하는 것은 유용하다.

그러나 현재는 많은 패션 쇼핑몰 회사가 독립적으로 데이터를 처리하고 분석하는 데에 한계가 있어 구체적인 제품의 상태나 스타일을 파악하기는 어렵다. M사의 플랫폼에서 제공되는 리뷰와 구매 만족도 점수 시스템만으로는 많은 리뷰를 보고 파악하는 데에 한계점이 존재하며, 높은 점수와 다르게 텍스트에 단점이 수록된 경우가 존재한다는 점, 매칭 가능한 스타일 탐색에 어려움이 존재한다는 점에서 개선이 필요하다.

1.2 연구 배경

영화 리뷰 요약 분석을 진행한 선행연구[1]에서는 온라인 사용자가 매일 수천 개의 영화 리뷰를 게시하기 때문에 리뷰를 수동적으로 요약하기는 어렵다고 판단하여 자동으로 긴 영화 리뷰를 요약해주는 연구를 진행하였다. 영화 리뷰에서 특징을 추출하고 이를 벡터 공간 모델 또는 특징 벡터로 표현한 후 나이브 베이즈 머신러닝 알고리즘을 사용하여 긍정과 부정으로 리뷰를 분류했다. 이후 가중 그래프 기반 알고리즘을 적용하여 각 리뷰 문장에 대한 순위 점수를 계산해 높은 순위 점수를 기준으로 선택하여 추출 요약을 진행했다. 본 연구에서는 딥러닝 사전학습 모델을 활용하고 모델에서 새로운 텍스트를 생성해내야 하기에 말이 되지 않는 표현이 만들어질 가능성이 존재하나 좀 더 유연한 접근이 가능한 생성 요약을 진행하여 딥러닝 기반 생성 요약의 우수성을 입증하고자 한다.

음식점 리뷰 감정 분석을 진행한 선행 연구[2]에서는 한국어로 작성된 음식점 리뷰를 대상으로, 감성분석을 수행하여 평가 항목별로 세분화된 평점을 제공 가능한 예측 방법론을 제안했다. 이를 위해, 음식점의 주요 평가항목으로 '음식', '가격', '서비스', '분위기'를 선정하고, 평가항목별 맞춤형 감성사전을 구축했다. 또한 평가항목별 리뷰 문장을 분류하고 감성분석을 통해 세분화된 평점을 예측하여 소비자가 의사결정에 활용 가능한 추가적인 정보를 제공했다. 이러한 선행연구를 바탕으로, 본 연구에서는 패션 도메인에 적용시켜 발전시키고자 한다.

---

[1] Khan, A., Gul, M.A., Zareei, M., Rajesh, R.B., Zeb, A., Naeem, M., Saeed, Y., & Salim, N. (2020). Movie Review Summarization Using Supervised Learning and Graph-Based Ranking Algorithm. *Computational Intelligence and Neuroscience, 2020*.

[2] SoJinSoo and Shin, Pan-Seop. (2020). Rating Prediction by Evaluation Item through Sentiment Analysis of Restaurant Review. Journal of The Korea Society of Computer and Information, 25(6), 81-89.

이외에도 패션 스타일을 군집화하는 방법을 제시한 선행 연구[3]에서는 패션 이미지 데이터셋에 대해 centroid 기반 밀도 중심 클러스터링 알고리즘을 통하여 패션 스타일을 군집화해 표현한다. 스타일 종류로는 Preppy, Goth, Hipster, Bohemian 등이 있으며, Preppy x Goth와 같이 두 가지 스타일을 믹스한 스타일로도 표현한다는 점에 있어 강점을 가진다. 패션 이미지를 다양한 스타일로 군집화한다는 면에 있어서 본 연구에서 원하는 태스크와 일치하지만, 해당 모델은 패션쇼 등의 이미지를 사용해 클러스터를 구성하였기 때문에 우리가 분석하려는 리뷰 이미지와는 거리가 있는 데이터셋이며, style 속성 이외에도 texture, fabric, part 속성을 함께 고려해 클러스터를 구성한다는 점에 있어 차이를 보인다.

패션 이미지의 색상을 분류하는 태스크에 대한 선행 연구[4]에서는 K-Means 클러스터링을 통해 이미지 내 옷의 색상을 구분한다. 이미지에서 MaskRCNN 모델을 이용해 이미지 내 옷 부분의 색상만을 추출하고, 한 이미지 내에서 옷 색상을 5가지 등으로 세부적으로 분류할 수 있는 모델을 구축하였다. 그러나 학습과 테스트에 사용된 이미지가 상의 또는 하의 중 하나에 해당되는 옷 하나에 대한 이미지를 이용했다는 점에 있어, 전신 리뷰 이미지를 분석하려는 우리의 태스크에는 정확히 일치하지 않으며, 한 이미지 내에서 상의와 하의를 각각 하나의 색상으로 나타내야 하는 모델을 구축할 필요가 있을 것으로 보인다.

2. 연구 목적 및 내용

2.1 연구 목적

[3] Chen, J., Yuan, H., Fang, F., Peng, T., & Hu, X.R. (2023). Unsupervised Fashion Style Learning by Solving Fashion Jigsaw Puzzles. 2023 IEEE International Conference on Multimedia and Expo (ICME), 1847-1852.

[4] 장혜림, 손봉기, 허권, 이재호. K-Means 클러스터링 기반 패션 이미지 색상 분류 구현. 한국통신학회 학술대회논문집.

본 연구는 패션 플랫폼 산업 전반에 대한 리뷰 분석을 기반으로 고객에게 가장 적합한 소비를 할 수 있도록 정보를 제공하는 것이 주된 목적이다. 구체적으로 Python언어와 Pytorch 프레임워크 등으로 딥러닝 사전 학습 모델들을 활용하여 fine-tuning을 거쳐 다양한 리뷰 분석 딥러닝 모델을 개발한다. 이는 소비자의 구매 실패율을 줄이고, 구매 시간 단축으로 소비자들의 만족도를 높이며 판매자에게도 매출 상승 등 긍정적인 영향을 미쳐 패션 플랫폼 시장에서 유의미한 결과를 이끌어낼 것으로 기대한다.

2.2 연구 내용

본 연구는 패션 플랫폼 M사에서 사용자들의 텍스트와 이미지 형태의 리뷰 데이터를 분석하는 것을 중심으로 진행한다. 해당 데이터는 사용자의 반응을 직접적으로 보여주므로, 해당 제품에 대한 사용자의 의견을 파악하는 데 유용하다.

수집된 데이터를 바탕으로 네 가지 리뷰 분석 자동화 모델을 개발한다. 첫 번째, 텍스트 리뷰 데이터의 자주 언급되는 색상, 사이즈 등의 단어 키워드를 분석한 후 키워드별 리뷰 요약을 제공한다. 두 번째, 텍스트 리뷰 데이터를 통해 해당 제품의 긍정 또는 부정의 정도를 표현한다. 세 번째, 이미지 리뷰 데이터로 사용자의 착용 스타일을 클러스터링한다. 마지막으로, 이미지 리뷰 데이터에서 해당 제품과 함께 매치한 옷이나 제품을 탐지하여 옷이나 제품의 색상을 HEX값을 기반으로 분류한다. 이 네 가지 분석 모델을 종합하여 소비자가 제품 리뷰를 편리하고 효과적으로 파악하는 데 도움을 준다.

나아가, 본 연구는 실제 서비스 플랫폼 상에서 적용 가능한 웹 User Interface 예시를 제한함으로써 다른 제품이나 플랫폼으로의 확장을 위한 인사이트를 제공한다.

3. 연구 방법론

3.1 텍스트 기반 분석

감정 분석(Sentiment Analysis)

감정 분석이란 텍스트에 들어있는 의견이나 감성, 평가, 태도 등의 주관적인 정보를 컴퓨터를 통해 분석하는 과정이다. 감정 분석은 최근 소셜 미디어를 포함한 온라인 플랫폼에서 블로그, 리뷰 등의 사용자 생성 콘텐트가 풍부해지면서 인기를 얻고 있으며, 이는 크게 두 가지 단계로 이루어져 있다. 첫 번째 단계로는 문서의 어떤 부분에 의견이 담겨있는지를 정의하고, 다음 단계에서는 이를 바탕으로 요약한다. 텍스트 내의 감정을 분류하거나 긍정과 부정의 정도를 점수화한다.

본 분석에서 사용할 모델은 KLUE-BERT, KLUE-RoBERTa, ALBERT-kor, KcBERT, KcELECTRAT이며, 모두 base 모델을 사용하였다. BERT(Bidirectional Encoder Representations from Transformer)는 Transformer의 encoder 부분만 사용한 모델로 Masked Language Model (MLM)과 Next Sentence Prediction (NSP) 방법을 채택하였고, RoBERTa는 BERT 모델이 충분하게 학습되지 않았다고 판단하여 정적 마스킹이 아닌 동적 마스킹을 사용하고 학습 시 NSP 작업을 사용하지 않고 더 많은 데이터를 사용하는 등의 변화를 통해 성능을 향상시켰다. 또한, KLUE(Korean Language Understanding Evaluation) 벤치마크는 2021년 Naver, Kakao, KIST등이 함께 만든 데이터셋으로, KLUE-BERT, KLUE-RoBERTa는 해당 데이터셋으로 학습시킨 모델이다. ALBERT-kor는, BERT 모델에서 factorized embedding layer parameterization, cross-layer parameter sharing의 방법으로 파라미터 수를 줄여 기존 BERT를 경량화시킨 한국어 버전의 모델이다. KcBERT는 공개된 다른 한국어 BERT 모델이 대부분 한국어 위키, 뉴스 기사, 책 등의 잘 정제된 기반으로 학습하여 정제되지 않은 댓글형 데이터셋에 적용하기 위하여 기존 모델과는 달리 온라인 뉴스에서 댓글과 대댓글을 수집해, 토그나이저와 BERT모델을 처음부터 학습한 사전학습된 BERT 모델이다. 비슷한 방법으로 KcELECTRA는 KcELECTRA보다 더 많은 데이터셋, 그리고 더 큰 General vocab을 통해 KcBERT 대비 대부분의 태스크에서 성능을 올린 모델이다.

평가 지표로는 분류 성능 평가 지표인 정확도, 재현율(Recall), 정밀도(Precision), F1-스코어(F1-Score)를 사용한다. 정확도는 실제 데이터가 예측 데이터와 얼마나 같은지를 판단하는 지표이고, 재현율은 실제 값이 사실인 대상 중 예측을 사실로 일치한 데이터의 비율을 나타낸다. 정밀도는 예측을 사실로 한 대상 중 실제로 사실인 데이터의 비율을 나타내며, 재현율이 높아지면 정밀도는 낮아지고

재현율이 낮아지면 정밀도는 높아지는 관계를 가지고 있다. 따라서 F1스코어는 정밀도와 재현율의 관계 확인이 가능하며 어느 한 쪽으로 치우지지 않는 수치를 나타낼 경우 상대적으로 높은 값을 갖는다.

텍스트 요약(Text Summarization)

텍스트 요약(Text Summarization)은 상대적으로 큰 원문을 핵심 내용만 간추려서 작은 요약문으로 변환한다. 일반적으로 텍스트 요약은 추출적 요약(Extractive Summarization)과 생성적 요약(Abstractive Summarization)으로 구분된다. 추출적 요약은 원본 내의 문장만을 활용하여 요약하는 방법이며, 생성적 요약은 유의어 등을 사용해 원본 외의 단어를 사용하여 요약하는 작업으로 추출적 요약보다 모델에서 새로운 텍스트를 생성해내야 하기에 말이 되지 않는 표현이 만들어질 가능성이 존재하나 좀 더 유연한 접근이 가능하다. 본 연구에서는 두 가지 방법을 적용해 본 후, 더 나은 성능을 보이는 방법을 채택하여 진행할 예정이다.

본 분석에서 사용할 모델은 추출적 요약으로는 감정 분석에서 사용한 모델과 동일한 KcELECTRA, KLUE-RoBERTa를 사용할 예정이며, 생성적 요약으로는 KoBART, KoT5를 사용할 예정이다. BART는 표준 transformer 기반 신경망 구조로, BERT와 GPT를 일반화한 것이라고 볼 수 있으며, BART의 핵심적인 장점은 noising의 유연성이다. 또한, 사전학습은 noise function으로 손상된 텍스트를 복구하도록 모델을 학습하는 방법으로 이루어지며, KoBART는 SKT에서 공개한 한국어 기반 BART 모델이다. T5는 모든 텍스트 기반 언어 문제를 text-to-text 형식으로 변환하는 통합 프레임워크를 도입한 모델이다. 입력 토큰에 self-attention 계산 시 offset boundary 내의 토큰들에 relative position encoding 값을 주었으며 텍스트를 모델에 대한 입력으로 사용하여 일부 대상 텍스트를 생성하도록 훈련시켰다. KoT5는 사전학습을 위해 한국어 위키 백과 및 신문기사 등이 사용된 한국어 버전의 T5 모델이다.

텍스트 요약에서의 평가 지표로는 요약본의 일정 부분을 비교하는 지표인 ROUGE (Recall-Oriented Understudy for Gisting Evaluation)를 이용한다. ROUGE는 성능을 평가하는 방법에 따라 ROUGE-1,-2, ROUGE-N, ROUG-L, ROUGE-S와 같이 여러 종류의 지표로 나뉜다. 예를 들어, ROUGE-N unigram, bigram, trigram 등 문장 간 중복되는 n-gram을 비교하는 지표이다. 본 연구에서 사용하는 태스크에 가장 적합한 ROUGE 지표를 찾아 성능을 평가한다.

3.2 이미지 기반 분석

객체 탐지(Object Detection)

객체 탐지란 컴퓨터 비전(Computer Vision)분야의 중요한 태스크 중 하나이며, 이미지 혹은 영상 내에서 찾고자하는 유의미한 객체를 찾아내는 작업을 말한다. 이미지 내에서 어떤 객체가 탐지되는지, 탐지된 특정 개체의 개수는 얼마나 많은지, 나아가서 얼굴 인식이나 비디오 추적을 하는 데에도 활용한다.

객체 탐지 과정은 두 가지 스텝으로 나누어볼 수 있다. 첫 번째는 Regional Proposal로 객체의 위치를 찾는 Localization과정이다. 바운딩 박스를 통해 객체의 위치를 파악하는 스텝이다. 이때 바운딩 박스란 객체가 존재하는 위치를 직사각형으로 표시한 것으로 x, y좌표로 구성된 데이터이다. 두 번째는 Classification으로 찾아낸 객체가 어떤 라벨에 해당하는지 분류하는 과정이다. 이 두 가지 스텝을 통해 어떤 객체가 어느 위치에서 탐지되었는지에 대해 알아낼 수 있다.

두 가지 스텝을 어떻게 수행하냐에 따라 객체 탐지 모델은 1-stage detector, 2-stage detector 두 가지 종류로 나누어볼 수 있다. 2-stage detector란 Regional Proposal과 Classification 과정을 순차적으로 진행하는 것을 말한다. R-CNN, Fast R-CNN, Faster R-CNN, Mask R-CNN 등 R-CNN계열 모델이 이에 해당한다. 두 가지 과정을 하나씩 진행하기 때문에 정확도는 높은 반면, 시간이 오래 걸린다는 단점이 있다. 반면, 1-stage detector란 두 과정을 동시에 진행하는 것을 말한다. SSD, RetinaNet, RefineDet, YOLO 등의 모델이 이에 해당하며, 정확도는 다소 낮을 수 있지만, 빠르다는 장점이 있다.

객체 탐지의 주요 평가 지표로는 앞서 언급한 분류성능평가지표인 Precision, Recall과 mAP가 있다. mAP란 mean Average Precision이다. 이해를 위해 AP 곡선(Average Precision Curve)에 대해 먼저 설명하자면, 이는 정밀도와 재현율을 고려한 종합 평가 지표를 말한다. 실제 AP값은 0에서 1 사이값으로 이루어지며, 1에 가까울수록 더 정확하다고 표현할 수 있다. 따라서 AP 값의 평균을 구한 값을 mAP라고 한다.

본 연구에서는 객체 탐지 모델로 Ultralytics의 YOLO(You Only Look Once)를 사용한다. YOLO는 CNN 기반의 객체 탐지 모형으로, 현재까지도 객체 탐지 관련 태스크의 state-of-the-art 모형으로 알려져 있다. YOLO는 1-stage detector 형태로 빠르다는 장점을 가지는 동시에, 다른 모델과 비교했을 때도 경쟁력 있는 성능을 보여준다. YOLO는 2016년 v1으로 시작하여, 2023년 1월에 출시된 v8까지의 버전이 존재한다. 본 연구에서는 현재 기준 최신 버전인 YOLOv8을 사용할 예정이다. 모델의 파라미터의 수에 따라 YOLOv8n, YOLOv8s, YOLOv8m, YOLOv8l, YOLOv8x 종류가 존재하고, n이 가장 적은 크기의 파라미터를 가지며 속도가 빠르다. 반면, x일수록 파라미터의 크기가 커지므로 성능이 좋다는 특징을 가진다. 본 연구에서는 여러 가지 모델을 사용해 최적의 성능을 구할 예정이다.

4. 연구 데이터

패션 쇼핑 플랫폼 사이트 M사에서 크롤링하는 방식으로 데이터를 수집하였다. 이 과정에서는 Python 언어와 BeautifulSoup 라이브러리 등을 활용하여 웹 스크래핑을 진행하였다. 업로드된 회원 후기 중 분석에 사용할 카테고리는 셔츠, 니트/스웨터, 맨투맨, 후드티셔츠, 긴소매 티셔츠, 반소매 티셔츠, 스커트, 데님팬츠, 코튼팬츠, 트레이닝/조거팬츠, 슬랙스, 숏팬츠, 롱 코트, 재킷, 블레이저, 숏패딩/패딩베스트, 후드집업, 카디건으로 총 18가지이다. 이 과정에서 수집한 정보는 리뷰 텍스트 데이터와 이미지 데이터, 기존 구매 만족도 점수 데이터이다.

4.1 텍스트 리뷰 데이터

텍스트 리뷰 데이터는 M사에 업로드된 회원 후기 중 '스타일 후기, '상품 후기', '일반 후기'를 크롤링 해 사용하였다. 웹 스크랩핑을 할 때에 각 카테고리별 랜덤하게 10여 종 정도의 상품을 선택하여 텍스트 리뷰 전체와 구매 만족도 점수(Rate)를 함께 수집하였다. 이 때, 구매 만족도 점수가 1~3인 것은 부정(label:0)으로 판단하고 4~5인 것은 긍정(label:1)로 판단하여 라벨 값을 설정하였다.

	A	B	C
1	Rate	Review	label
2	1	무슨 2군데나 터진 불량품을 보내나요? 검수 안하나요? 너무나도 어이없어서 교환 환불 하고싶은데 택배 보낼 시간이 없어서 튀어나온 털만 그냥 가위로	0
3	5	생각보다 두껍고 무겁네요 마침 날씨도 쌀쌀해져서 딱인듯	1
4	4	내구성이 상당히 좋은 니트라고 생각됩니다. 목부분, 손목, 허리 부분 시보리가 상당히 튼튼하네요다만 베레모 55사이즈를 착용하던 제가 머리 넣기가 불	1
5	1	마지막으로 믿고 유튜버 콜라보 샀는데 진짜 깡형은 믿거할예정 세탁 후 줄어듬이 젤 큰듯	0
6	5	입었을 때 중량이 생각보다 무겁습니다두껍기도 두꺼워서 단품으로 입거나 오버사이즈의 아우터를 입어야 할 것 같네요와이드한 슬렉스, 딥한 색상(생지,	1
7	1	깡 콜라보 하자마자 블루 칼라 먼저 구입.2번이나 사이즈가 잘못와서 3번만에 제대로된 사이즈를 받음. 몇번이나 잘못된 옷이 온거는 진짜 빡쳤지만 CS팀	0
8	3	L로 했다가 부해보여서 바로 M으로 교환하여 한 2주만에 드디어 입어보네요 같은 부산지역이라 빨리 올 줄 알았는데 웬 대전으로 갔다가 다시 내려오던ㄷ	0
9	5	뒷면에 약간 오염된 부분이 있었지만 이미.. 택을 뜯은 상태여서 ㅠㅠ 포기하고 입으려구요 100~105사이즈 입는데 괜찮아요	1
10	5	제가 원하던 핏이여서 조은데 먼지가 너무잘묻어요	1
11	5	니트 생깍보다 두꺼워서 봄에는 입기 힘들것 같네요	1
12	2	생각보다는 영 핏이 좋진 않네요 그냥 기본템 좋지도 안좋지도않은	0
13	1	상품은 대만족하지만 콘텐츠 이후 물량 부족과 배송지연 이거에 대한 씨에스는 엉망 임	0
14	3	핏 모두 마음에드는데 박음질이 이상하고 올이 약간 나가있는 곳이 세 군데 정도 있어서 조금 아쉬워요. 교환하면 너무 오래걸릴 것 같아서 그냥 입으려ㄱ	0
15	5	내부 마감이 진짜 안 좋은데 핏은 맘에 들어서 입어요요	1
16	5	소매부분이랑 목부분이 좀 타이트하네요 기장이랑 색깔은 이뻐요	1

[그림1] 수집한 텍스트 데이터 예시

수집한 데이터를 확인해보면 기존 구매 만족도 점수가 긍정적인 쪽에 분포되어 있는 것을 알 수 있다. 심지어 텍스트에는 부정적인 내용이 내포되어 있음에도 구매 만족도 값은 긍정으로 매겨져 있는 경우가 다수 존재한다. 이는 구매자들이 상품에 대한 만족스럽지 못한 부분이 있더라도 리뷰를 달아 혜택을 얻기 위해 깊은 고민 없이 합리적이지 않은 점수를 매겼을 것이라 판단한다. 또한, 긍정적인 리뷰와 부정적인 리뷰의 큰 불균형은 추후의 분석 결과에 방해를 받기 때문에 불균형 완화의 필요성도 있다고 판단하였다. 따라서 본 연구에서는 기존 구매 만족도를 재조정하는 방안을 채택하였다.

기존 구매 만족도 재조정은 GPT-4의 API를 사용하여 진행한다. 이 API는 3가지의 role은 system, user, assistant로 나누어져 있고, 본 연구 과정에서는 system과 user만을 사용하였다. role의 system을 통해 원하는 태스크에 간단한 설명을 함께 입력하여 주는 incontext-learning을 수행할 수 있으며, 본 연구에서는 “너는 패션 옷 리뷰에 담긴 고객 감정을 분석하고 탐지하는 AI 언어모델이야"라고 입력하였고, user 부분에는 옷 리뷰를 분석하여 각 고객별 감정이 긍정인지 부정인지 구분하도록 질문을 입력시켰다.

```
def analyze_review(review):

  try:
    messages = [
            {"role": "system", "content": "너는 패션 옷 리뷰에 담긴 고객 감정을 분석하고 탐지하는 AI 언어모델이야"},
            {"role": "user", "content": f"다음 옷 리뷰를 분석하여 각 고객 별  감정이 긍정인지 부정인지 판단해 알려줘. 대답은 다른 추가적인 설명없이 '긍정' 또는 '부정'  둘 중 하나의 단어로 대답해야 해: {review}"}
        ]

    completion = openai.ChatCompletion.create(
            model="gpt-4",
            messages = messages,
            max_tokens=5,
            n= 1,
            stop=None,
            temperature=0.5
        )

    response= completion.choices[0].message.content
    print(response)
    return response
```

[그림2] 라벨값 재조정 과정

위와 같이 GPT-4를 사용하여 라벨 값을 재조정(Relabel)한 후, 긍정적인 리뷰가 부정적인 리뷰에 비해 개수가 훨씬 많기 때문에 데이터의 불균형을 완화하기 위하여 긍정과 부정 리뷰의 비율을 약 6.5 : 3.5로 설정하여 해당 긍정적인 리뷰의 개수를 그 비율에 맞게 랜덤하게 삭제하는 작업을 진행하였다. 최종적으로 약 12,000개의 학습 데이터셋을 구축하였다. 본 연구에서는 GPT-4가 제시해준 라벨 값과 팀원들의 판단을 비교하여 GPT-4가 제시해준 라벨의 정확성을 입증하였다.

	A	B	C	D
1	Rate	Review	label	Relabel
2	1	무슨 2군데나 터진 불량품을 보내나요? 검수 안하나요? 너무나도 어이없어서 교환 환불 하고싶은데 택	0	부정
3	5	맨살에 입어도 부드럽고 재질은 되게 좋아요두께감이 두꺼워 초겨울 쯤 돼야 입을 거 같습니다 기장도	1	긍정
4	5	진짜 완전 부드럽고 핏 너무 이쁘게 나와요.요즘 날씨에 실내에선 단품, 밖에선 이너로 입는데 너무 좋너	1	긍정
5	5	깡스타일리스트 콜라보는 항상 믿을만 합니다. 옷 진짜 두껍고 색깔도 예쁘고 착용감도 좋아요	1	긍정
6	5	생각보다 두껍고 무겁네요 마침 날씨도 쌀쌀해져서 딱인듯	1	부정
7	5	색상은 알아듣기 쉽게 딸기우유, 오피스텔 상가 분양할 때 나눠주는 분홍행주 컬러예요. 운동 좋아하고	1	긍정
8	5	옷도 예쁘고 콜라보제품이라 저렴하게 잘 산거같습니다 ㅎㅎ	1	긍정
9	5	너무이뻐용 다릉색도 빨리내주세용 다살거같아요	1	긍정
10	4	내구성이 상당히 좋은 니트라고 생각됩니다. 목부분, 손목, 허리 부분 시보리가 상당히 튼튼하네요다만	1	부정
11	5	무난하게 입을수 있을것 같아요두께감이 있어요	1	긍정
12	1	마지막으로 믿고 유튜버 콜라보 샀는데 진짜 깡형은 믿거할예정 세탁 후 줄어듬이 젤 큰듯	0	부정
13	5	입었을 때 중량이 생각보다 무겁습니다두껍기도 두꺼워서 단품으로 입거나 오버사이즈의 아우터를 입	1	부정
14	5	USJ 닌텐도 가는데 친구랑 마리오 루이즈 맞출려고 구매했습니다. 쨍한 초록색이 예쁠까 생각했는데 예	1	긍정
15	1	깡 콜라보 하자마자 블루 칼라 먼저 구입.2번이나 사이즈가 잘못와서 3번만에 제대로된 사이즈를 받음.	0	부정
16	4	품은 넓은데 기장은 좀 애매하게 짧아요. 그래도 좋은 소재 좋은 가격이라 만족해요.	1	긍정

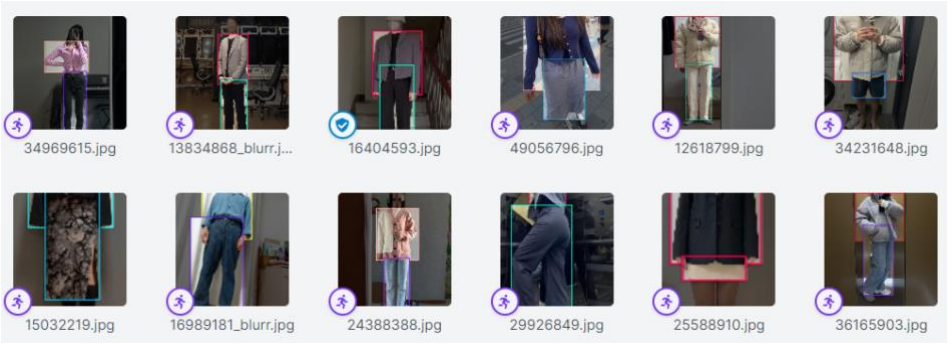
[그림3] 라벨값 재조정 결과 예시

4.2 이미지 리뷰 데이터

이미지 리뷰 데이터는 M사에 업로드된 회원 후기 중 전신이 나온 이미지 형태의 '스타일 후기'를 크롤링해 사용하였다. 웹 스크랩핑을 할 때에 각 카테고리에 해당되는 상품군 중 판매순으로 정렬하여 랜덤하게 약 30-40여 종 정도의 상품을 선택하였다. 이후 상품 각각에 대해 '스타일 후기'란의 1,2페이지에 해당하는 이미지 약 20여 장을 가져왔다. 따라서 한 카테고리별 약 600-700장의 이미지가 수집되었다.

[그림4] 얼굴 모자이크 처리한 이미지 예시

전체 수집한 이미지 데이터셋 중 사용자의 얼굴이 나온 데이터가 상당 수 존재했다. 초상권 보호를 위해 얼굴이 포함된 이미지에 대해 블러미(blur.me)를 이용해 얼굴 모자이크 처리를 진행하였다. 블러미는 AI를 이용해 얼굴의 위치를 인식하고, 이에 대해 자동으로 모자이크 처리해주는 웹 프로그램이다.



[그림5] 로보플로우를 이용한 라벨링 예시

모자이크 처리 후, [그림 5]와 같이 전체 이미지 데이터셋에 대해 로보플로우(Roboflow)를 이용해 바운딩박스를 생성하는 라벨링을 진행하였다. 로보플로우란 라벨링 툴로써 이미지 데이터에 대해 어노테이션 및 데이터 증강, 모델의 성능을 측정할 수 있는 플랫폼이다. 라벨값은 18가지의 의류 카테고리로, [‘blazer’, ‘cardigan’, ‘coat’, ‘cottonpants’, ‘denimpants’, ‘hoodies’, ‘jacket’, ‘longsleeve’, ‘mtm’, ‘padding’, ‘shirt’, ‘shortpants’, ‘shortsleeve’, ‘skirt’, ‘slacks’, ‘sweater’, ‘trainingpants’, ‘zipup’]이 이에 해당한다.

[그림6] 이미지 증강 결과 예시

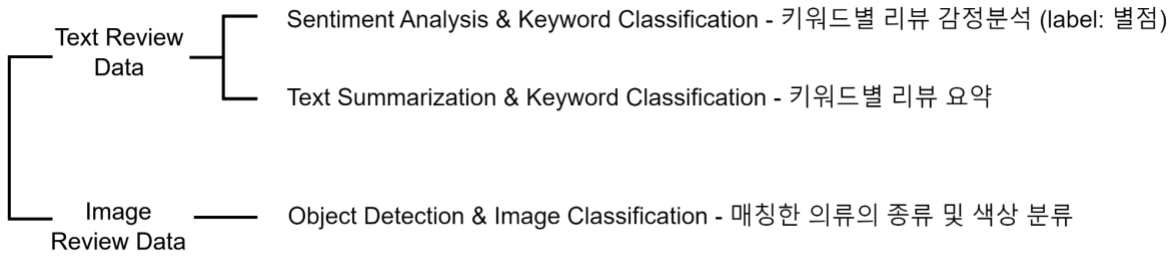
라벨링 후 데이터셋에 대해 이미지 증강(augmentation)을 완료했다. 이를 위해 우선 학습데이터셋:검증데이터셋:평가데이터셋의 비율을 8:1:1로 맞추고, 학습데이터셋에 대해 3배의 양으로 데이터를 증강했다. 이때 증강 기법으로는 최대 20% 이미지 확대, 최대 15° 회전, 이미지 내 25% 정도 그레이 스케일 처리, -25%에서 25%사이의 밝기 조절, -25%에서 25%사이의 노출 조절, 최대 5%의 픽셀에 대해 노이즈값 부여 기법을 사용하였다.

이렇게 전처리된 데이터를 최종 데이터셋으로 사용했으며, 최종 사용한 이미지 데이터의 수는 14,972장이다.

5. 연구 프로세스 및 결과

	System Architecture
Data	MUSINSA text & image
Database	MYSQL
Language(framework)	Python(Pytorch, Tensorflow)
Web	HTML, CSS, javascript

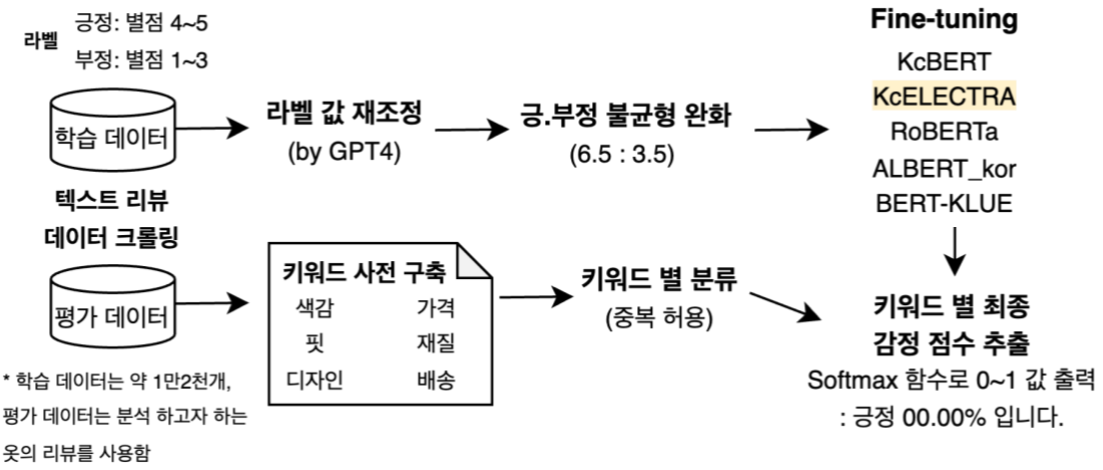
[표1] 시스템 아키텍처



[표2] 방법론 요약

본 연구에서 활용하는 시스템 아키텍처와 방법론을 요약하면 [표1]과 [표2]와 같다. 현재까지는 데이터 수집 및 전처리, 키워드별 리뷰 감정 분석, 객체 탐지, 색상 분류, 그리고 간단한 웹 페이지 개발을 완료하였다.

5.1 텍스트 리뷰 감정 분석



[구조도1] 텍스트 리뷰 분석 프로세스

데이터 수집 단계에서 수집한 텍스트 리뷰 데이터와 최종 라벨 값을 바탕으로 텍스트 리뷰의 중요 키워별 감정을 요약하여 제시하기 위해 감정 분석을 진행하였다. 여기에서 키워드는 수집된 텍스트 리뷰 데이터에서 많이 언급된 키워드 6가지를 추출하여 키워드 사전을 구축하였고, 각 키워드에 해당되는

리뷰를 분류하였다. 이때 텍스트 리뷰 안에 두 가지 이상의 키워드에 내용을 포함하고 있다면, 두 가지 이상의 키워드 부분에 해당 텍스트 리뷰를 모두 분류하였다.

Fine-tuning

Fine tuning은 사전 학습된 모델의 가중치를 새로운 데이터에 맞게 세밀하게 조정하여 성능을 향상시키는 방법이다. 따라서 본 연구에서는 사전 학습된 모델에 연구에서 사용할 텍스트 리뷰 데이터셋을 fine-tuning하였다. fine-tuning을 진행한 사전 학습 모델은 ALBERT-kor, KcELECTRA, KcBERT, KLUE-RoBERTa, KLUE-BERT 총 5가지의 base 모델이다. max_length는 64로 설정하여 padding을 진행하였고, 여러 가지 하이퍼파라미터로 실험을 진행하였으며, 최종적으로 사용한 파라미터는 early stopping을 적용하여 batch_size 64, patience 10으로 설정하였다. 또한, 옵티마이저는 Adam, 손실 함수는 이중 분류이므로 binary cross entropy를 사용하였다.

위와 같은 방법으로 실험을 진행한 결과, KcELECTRA 모델이 여러 지표에서 성능이 가장 우수하였다. 따라서, 본 연구에서는 KcELECTRA base 모델을 채택하였다.

모델 *base	정확도	재현율	정밀도	f1-스코어
ALBERT_kor	0.9183	0.9176	0.9007	0.9082
KcELECTEA	0.9188	0.9159	0.9036	0.9092
KcBERT	0.9101	0.9103	0.8895	0.8984
KLUE-RoBERTa	0.9123	0.9109	0.8940	0.9015
KLUE-BERT	0.8912	0.8771	0.8880	0.8819

[표3] Fine-tuning 결과

키워드 사전 구축

감정 분석 모델을 구축한 후, 각 키워드별로 감정 점수를 제시하기 위해 텍스트 리뷰에서 명사를 추출하여 빈도수를 확인하였다. 그 다음 각 키워드별로 많이 언급되는 명사를 뽑아 키워드 사전을

구축하였다. 이를 통해, 새로운 옷에 대한 텍스트 리뷰들이 주어졌을 때, 해당 키워드의 리뷰 부분으로 분류가 가능하도록 한다.

<색감>
색감, 색상, 오트밀, 그레이, 색, 색도, 색깔, 블랙, 회색, 블루, 아이보리, 흰색, 컬러, 채도, 명도, 어두운, 파란색, 연한, 화이트, 멜란지, 검정, 연한

<핏>
사이즈. 핏, 기장, 길이, 품, 소매, 핏감, 루즈, 루즈핏, 널널, 스펙, 정사이즈, 라지, 활용, 체형, 키, 정핏, 실루엣, 와이드, 흐물흐물, 몸무게, 분위기, 넥라인, 넥부분,

<디자인>
디자인, 스타일, 심플, 트렌드, 트랜디, 연출, 코디, 스탠다드, 유행, 꾸안꾸룩, 데님, 네추럴, 트렌디, 무드, 휘뚜르, 마뚜르, 매치, 매칭

<가격>
가격, 가성, 할인, 가성비, 블프, 블랙프라이데이

<재질>
재질, 두께, 촉감, 따뜻, 퀄리티, 보풀, 소재, 원단, 니트, 집업, 겨울, 가을, 지퍼, 퓸질, 마감, 실밥, 밑단, 옷감, 까슬까슬, 부드러움, 관리, 세탁, 줄어듬, 먼지, 쫀쫀, 보온, 무게, 한겨울, 냄새, 건조기, 짜임새, 기모, 감촉, 싸구려

<배송>
배송, 포장, 교환, 딜레이, 검수, 불량, 불량품, 서비스, 주문, 반품

[그림7] 키워드 사전 예시

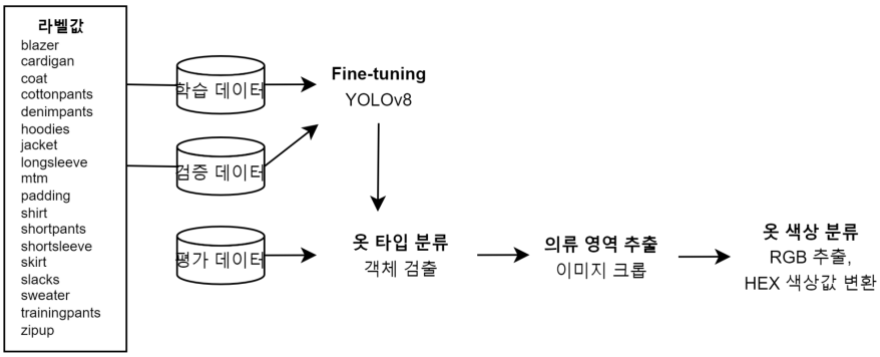
감정 점수 추출

앞서 진행한 KcELECTRA의 fine-tuning을 진행하여 얻은 가중치와 구축한 키워드 사전을 활용하여 최종적으로 요약하고자 하는 옷에 대한 키워드별 감정 점수를 추출한다. 먼저, 재조정한 부정, 긍정 라벨 값을 각각 0과 1로 지정한 다음, 원 핫 인코딩을 한 후 부정인 경우와 긍정인 경우의 확률 값을 sotfmax 함수를 활용하여 구하였다. 다시 말해, 리뷰 한 개당 부정과 긍정의 정도가 각각 몇 퍼센트의 비중을 차지하는지를 나타내었으며 각 확률의 합은 1이 된다. 다음으로, 각 리뷰의 부정과 긍정의 정도를 비교하여 더 확률 값이 높은 값을 택하여 해당 리뷰가 부정인지 긍정인지 판단하였고, 더 높은 확률 값을 부정과 긍정의 정도(퍼센트)로 활용하였다. 마지막으로, 하나의 옷마다 총 요약된 감정 점수를 제시하기 위하여 각 리뷰별로 앞에서 구한 확률 값을 모두 더하여 총 리뷰 개수로 나누어 최종적인 감정 점수를 추출하였다. 본 프로세스는 모든 키워드에서 동일하게 진행되었다.

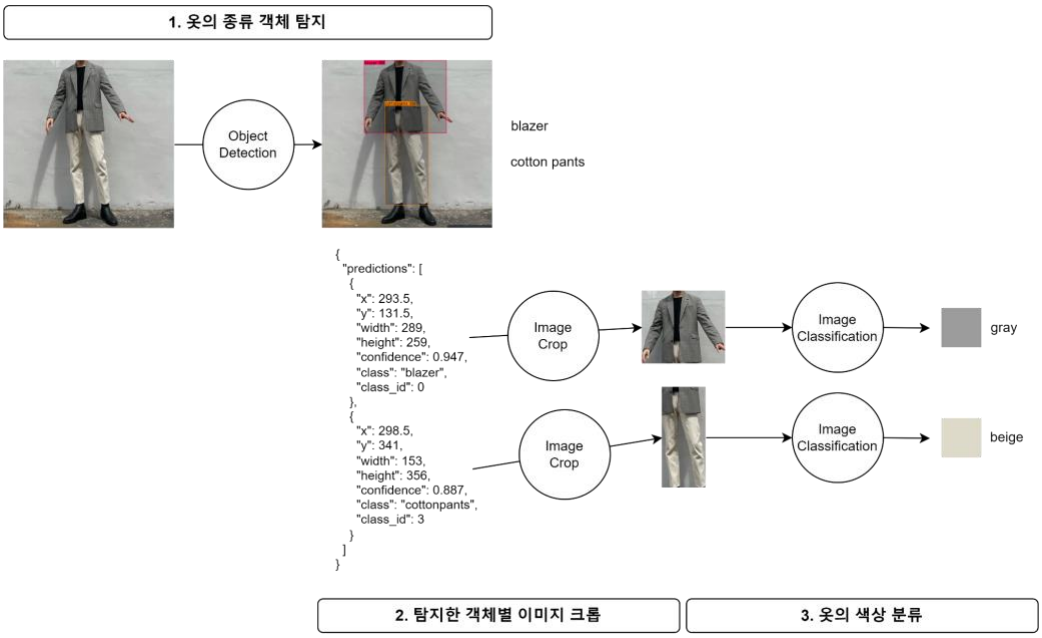
Keyword	Results
가격	75.88% 긍정입니다.
디자인	80.52% 긍정입니다.
배송	66.49% 긍정입니다.
색감	81.90% 긍정입니다.
재질	79.00% 긍정입니다.
핏	76.10% 긍정입니다.

[표4] 감정 점수 추출 예시

5.2 이미지의 의류 종류 및 색상 분류



[구조도2] 이미지 리뷰 분석 프로세스



[구조도3] 이미지 리뷰 분석 예시

이미지 리뷰 분석의 전체 프로세스는 [구조도 2]와 같으며, [구조도 3]은 한 이미지에 대해 전체 프로세스를 적용한 예시이다. 프로세스는 크게 3가지로 구성되는데 이를 아래에서 단계별로 자세히 설명하겠다.

의류 객체 검출

Class	Precision	Recall	mAP50	mAP50-95
all	0.837	0.623	0.834	0.736
blazer	0.931	0.792	0.919	0.804
cardigan	0.822	0.713	0.841	0.747
coat	0.747	0.833	0.862	0.785
cottonpants	0.887	0.688	0.859	0.767
denimpants	0.873	0.737	0.907	0.786
hoodies	0.711	0.457	0.682	0.587
jacket	0.789	0.504	0.738	0.655
longsleeve	0.854	0.544	0.802	0.703
mtm	0.878	0.6	0.799	0.712
padding	0.614	0.833	0.866	0.768
shirt	0.952	0.508	0.76	0.679
shortpants	0.919	0.75	0.874	0.712
shortsleeve	0.958	0.907	0.978	0.856
skirt	0.826	0.757	0.87	0.682
slacks	0.867	0.636	0.847	0.73
sweater	0.728	0.474	0.683	0.621
trainingpants	0.751	0.478	0.723	0.655
zipup	1	0	0.995	0.995

[그림8] 이미지 객체 검출 결과

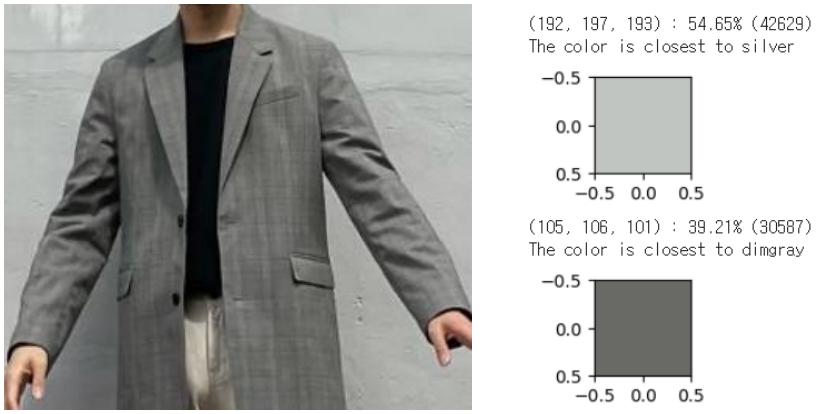
리뷰 이미지가 어떤 카테고리의 옷을 담고 있는지 알아보기 위해서 우선 YOLO 모델을 이용해 객체 검출을 진행하였다. 이때 YOLOv8n, YOLOv8m 모델 두 가지로 실험하였으며, 여러 가지 경우의 수로 하이퍼 파라미터값에 변화를 주어 실험하려고 한다. 현재까지 가장 좋은 성능을 낸 모델의 하이퍼 파라미터값은 50 epoch, 이미지 사이즈 640*640, patience 50, YOLOv8m 모델을 사용해 학습시킨 모델이다.

현재까지 이미지 객체 검출 결과의 성능은 [그림 8]과 같다. 모든 클래스에 대해 83%의 mAP값을 가지며, 각 클래스별 정확도는 68%에서 99%로 다양하다. 이중 정확도가 낮은 클래스에 대해 이미지를 추가하고, 이미지 증강을 여러 방법으로 시도해 정확도를 높일 예정이다.

검출된 객체별 이미지 크롭

객체 검출이 완료된 이미지는 탐지된 객체가 속하는 의류 카테고리 및 이에 대한 바운딩박스 영역이 데이터로 저장된다. 하나의 객체에 대한 색상을 추출하기 위해 객체별 이미지를 크롭(crop)하는 과정이 필요하다. 따라서 파이썬의 PIL 패키지를 이용해 저장된 객체의 바운딩박스 영역만큼 이미지를 크롭하는 간단한 코드를 구성했으며, 이후 모델에 해당 코드를 추가할 예정이다.

의류의 색상 추출

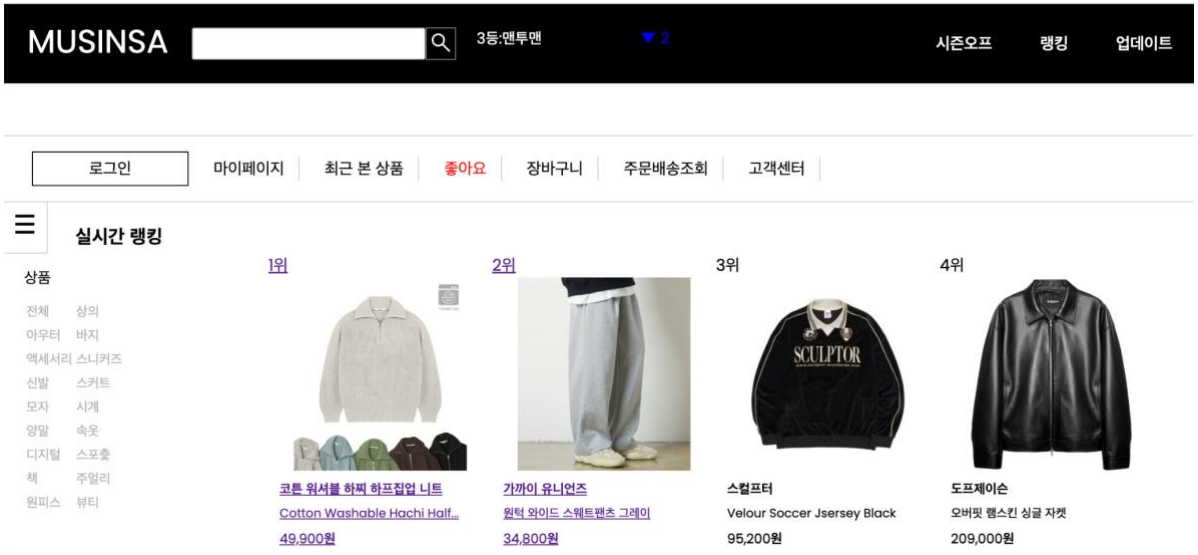


[그림9] 의류 색상 추출 예시

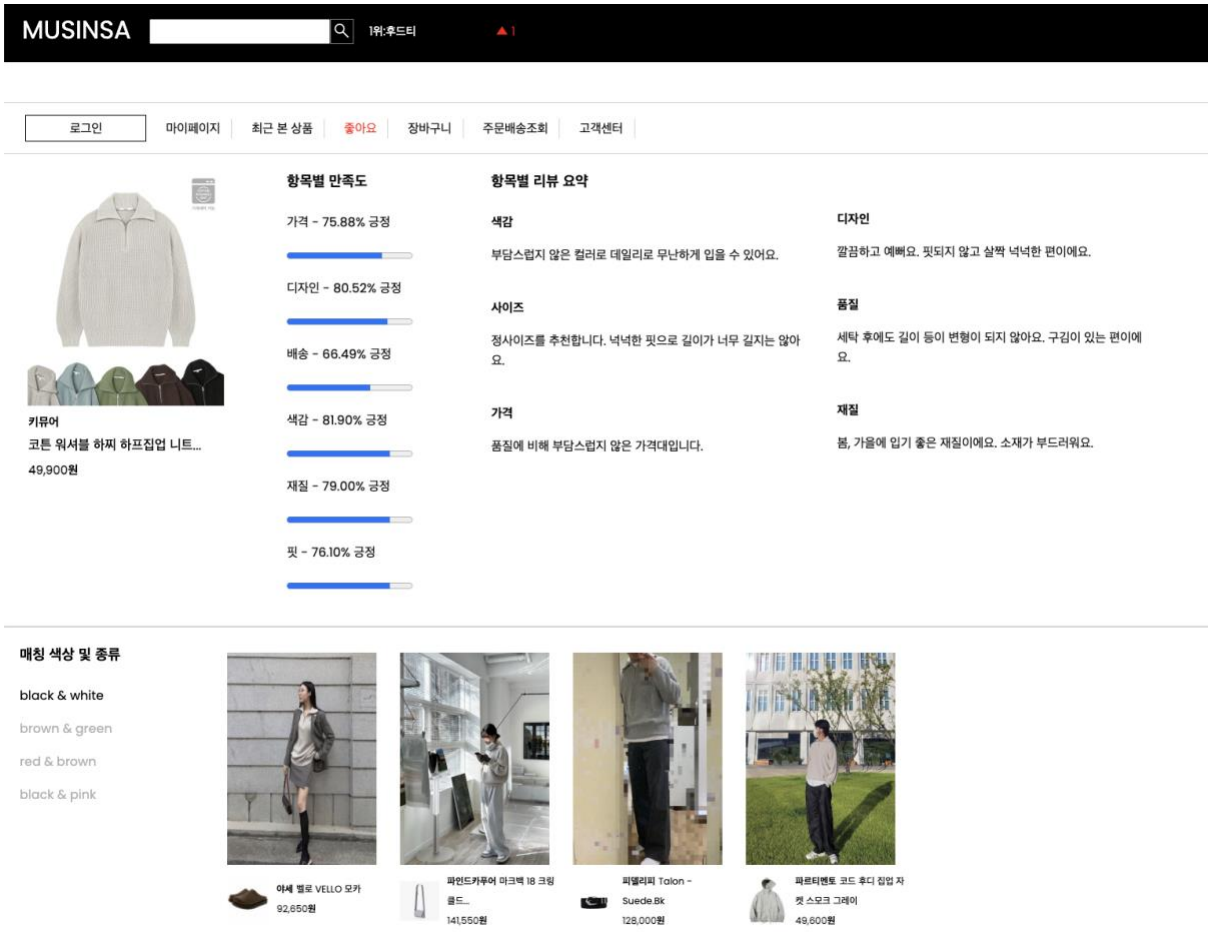
크롭한 이미지는 이제 이미지 내 하나의 의류를 나타낸다. 이를 가지고 해당 의류의 색상을 추출하는 코드를 구현하였다. 색상 추출은 파이썬의 색상을 추출하는 extcolors 패키지를 이용하여 픽셀의 RGB값을 추출한다. 이 RGB값을 HEX값으로 변환하고, webcolors 패키지의 CSS3_HEX_TO_NAMES.items()를 활용해 HEX값과 색상 이름이 나온 데이터 중 가장 비슷한 색으로 분류해, 그 값을 결과값으로 가지게 된다. [그림 9]의 예시를 보면, silver와 dimgray값이 결과값으로 나오는 형태임을 알 수 있다. 이 색상값을 데이터로 저장하고, 같은 색상끼리 이미지를 분류해 웹 페이지에 보여줄 계획이다.

5.3 웹 페이지 개발

본 연구에서는 제시한 프로세스의 활용 인사이트를 제공하기 위하여 웹 페이지를 통해 특정 상의와 하의 각 한 가지에 대해 결과 예시를 보여준다. 웹 페이지는 HTML, CSS, Javascript 언어로 개발을 진행하였으며, 최종적으로 각 옷에 대하여 항목별 만족도, 항목별 리뷰 요약, 매칭 색상 및 종류를 정리하여 보여주는 데에 그 목적이 있다. 나아가 이는 다른 제품이나 플랫폼으로의 확장 가능성을 제공한다.



[그림10] 웹 메인 페이지 페이지 개발 예시



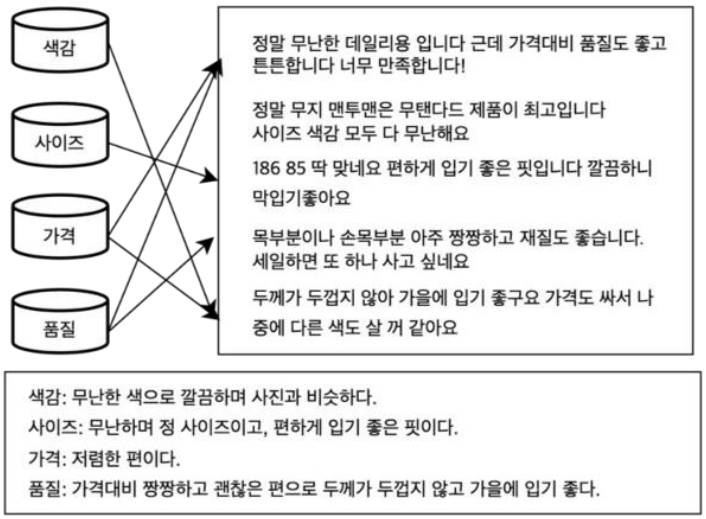
[그림11] 웹 리뷰 요약 페이지 개발 예시

6. 연구 결과의 중요성 및 기대효과

본 연구는 사용자 리뷰 활용 강화에 기여할 것으로 기대된다. 본 연구의 분석 방법 및 개발 모델은 해당 상품 뿐만 아니라 전체 상품, 그리고 M사의 플랫폼 뿐만 아니라 다른 플랫폼으로의 확장성을 가진다. 따라서 실무자가 본 연구의 결과를 바탕으로 효율적인 서비스를 제공할 수 있을 것으로 기대할 수 있다. 소비자 입장에서는 한눈에 보기 어려운 단순히 나열된 리뷰와 구매 만족도 점수만으로 구매 의사결정을 해야 한다는 한계점을 극복할 수 있다. 또한, 판매자 입장에서는 소비자의 구매 만족도와 활용 사례를 간편하면서도 구체적으로 파악할 수 있게 됨에 따라 소비자의 요구를 더욱 세밀하게 반영하는 상품 개발 전략을 세울 수 있게 된다. 이러한 과정은 소비자의 구매 만족도 상승, 판매자의 매출 상승과 패션 플랫폼 시장 전반에 긍정적인 영향을 미칠 것이라 기대된다.

7. 향후 계획

7.1 텍스트 리뷰 요약



[그림12] 텍스트 요약 결과 예시

앞서 진행한 감정 분석에 이어서 텍스트 리뷰의 요약을 진행하고자 한다. 방법론 파트에서 언급한 것처럼 KcELECTRA, KLUE-RoBERTa, KoBART, KoT5의 사전 학습 모델을 사용하여 추출적 요약(Extractive Summarization)과 생성적 요약(Abstractive Summarization) 두 가지 방법으로 비교하며 최적의 모델을 선정하여 텍스트 리뷰 요약 기능을 제공할 예정이다.

7.2 이미지 내 의류 객체 검출 및 색상 분류 모델

현재까지 진행한 이미지 내 의류의 종류를 구분하는 객체 검출 모델의 mAP값은 83% 정도에 불과하다. 따라서 검출 성능이 낮은 클래스의 이미지를 더 수집 및 전처리하고, 증강 과정을 거쳐 다양한 클래스에 대해서도 안정적인 성능을 갖는 모델을 만드는 것이 목표이다. 또한 한 모델을 학습시키는 데에 시간이 많이 걸려 다양한 하이퍼 파라미터를 시도하지 못했는데, 중간 보고 이후 이에 주력하여 더 많은 에포크 값과 하이퍼파라미터값에 변화를 주어 최적의 성능을 연구해볼 계획이다.

현재에는 이미지 검출과 이미지 크롭, 이미지 색상 분류 과정이 각각의 코드로 구성되어 있어 입력 데이터에서 출력 데이터를 얻기까지 여러 코드를 실행해야 한다는 단점이 있다. 실험 후 시간이 여유롭다면, 해당 코드를 하나로 합쳐 한 모델 안에서 동작하도록 만드는 것이 최종 목표이다.

7.3 웹 페이지 개발

앞서 제시한 웹 개발 예시 결과에 이어서 최종적으로 분석한 텍스트와 이미지 분석을 포함한 결과를 웹사이트에 적용하여 실제 새로운 옷에 대한 텍스트와 이미지 리뷰가 입력되었을 때 어떤 결과를 나타낼 수 있는지 확인할 수 있도록 웹사이트의 프론트엔드 부분을 개발 완료할 예정이다.

참고 문헌

Sizov, G. (2010). Extraction-Based Automatic Summarization: Theoretical and Empirical Investigation of Summarization Techniques.

Khan, A., Gul, M.A., Zareei, M., Rajesh, R.B., Zeb, A., Naeem, M., Saeed, Y., & Salim, N. (2020). Movie Review Summarization Using Supervised Learning and Graph-Based Ranking Algorithm. Computational Intelligence and Neuroscience, 2020.

Liu, Y., & Lapata, M. (2019). Text Summarization with Pretrained Encoders. ArXiv, abs/1908.08345.

Zhong, M., Liu, P., Chen, Y., Wang, D., Qiu, X., & Huang, X. (2020). Extractive Summarization as Text Matching. Annual Meeting of the Association for Computational Linguistics.

Xiao, H., Rasul, K., & Vollgraf, R. (2017). Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. ArXiv, abs/1708.07747.

Zhong, M., Liu, P., Chen, Y., Wang, D., Qiu, X., & Huang, X. (2020). Extractive Summarization as Text Matching. Annual Meeting of the Association for Computational Linguistics.

Chung, J. (2023). Proposal of Customer Experience-based Service Innovation Framework through Case Analysis of Domestic Unicorn Service Companies. Korea Institute of Design Research Society.

Lee, D., Jo, J.-C., & Lim, H.-S. (2017). User sentiment analysis on Amazon fashion product review using word embedding. Journal of the Korea Convergence Society, 8(4), 1–8. doi:10.15207/jkcs.2017.8.4.001

소진수 and 신판섭. (2020). 음식점 리뷰 감성분석을 통한 세부 평가항목별 평점 예측. 한국컴퓨터정보학회논문지, 25(6), 81-89.

[bert] keybert로 리뷰 키워드 추출하기. (n.d.). Retrieved from https://velog.io/@mare-solis/BERT-keyBert%EB%A1%9C-%ED%82%A4%EC%9B%8C%EB%93%9C-%EC%B6%94%EC%B6%9C%ED%95%98%EA%B8%B0

Text summarization. (n.d.). Retrieved from https://www.sciencedirect.com/topics/computer-science/text-summarization

Supersimples. (n.d.). Retrieved from https://github.com/supersimples/musinsaclone

(N.d.-a). Retrieved from https://m.ddaily.co.kr/page/view/2023042609504238764

Grootendorst, M. P. (n.d.). Retrieved from https://maartengr.github.io/KeyBERT/index.html

Lee, D., Jo, J.-C., & Lim, H.-S. (2017). User sentiment analysis on Amazon fashion product review using word embedding. Journal of the Korea Convergence Society, 8(4), 1–8. doi:10.15207/jkcs.2017.8.4.001

Wankhade, M., Rao, A.C., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review, 55, 5731 - 5780.

Wouter, V. G. (2020, May 25). SCAN: Learning to Classify Images without Labels. arXiv.org. https://arxiv.org/abs/2005.12320

Ester, M. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. https://www.semanticscholar.org/paper/A-Density-Based-Algorithm-for-Discovering-Clusters-Ester-Kriegel/5c8fe9a0412a078e30eb7e5eeb0068655b673e86

Deep Adaptive Image Clustering. (2017, October 1). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/8237888

https://github.com/zalandoresearch/fashion-mnist

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. ArXiv, abs/1909.11942.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv, abs/1907.11692.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. Annual Meeting of the Association for Computational Linguistics.