**[ 3차(6월 2주차) 프로젝트 수행 일지 ]**

| **프로젝트 타이틀** | 판매 데이터 분석을 통한 수요 예측 | | |
| --- | --- | --- | --- |
| **프로젝트 팀명** | 3조 - 데이터 맛집 | 프로젝트 팀원 | 황진웅(팀장),  김철기, 박건하, 성원선, 홍석우 |

※ 프로젝트 수행 일지는 매 주 금요일 17시까지 최종본을 저장해주세요

**1. 프로젝트 수행 계획 및 현황**

| **1. 프로젝트 수행 계획 및 현황** | | |
| --- | --- | --- |
|  | **금주 프로젝트 수행 내용** | **차주 프로젝트 수행 계획** |
|  | **1) 데이터 추가 수집**  **1.1) 데이터 불균형**  - 1~3점 데이터가 부족하여 11번가와 옥션 데이터 추가 수집 계획이었으나 오버샘플링으로 문제 해결하여 추가 수집 불필요  **1.2) 카카오맵 API이용**  - 마켓컬리의 상품 데이터 위주로 사용하여  수집 중지, 추후 필요하면 수집 예정  **1.3) 맞춤형 상품 추천 시스템**  - 마켓컬리의 밀키트 상품 데이터 수집  - 기상청API를 통해 8도와 제주도를 대표하는  도시들(22개)의 5년간 날씨 데이터 수집  - 네이버 데이터랩을 통해 성별과 연령별 음식  종류에 대한 검색량(합계) 수집(2024~2025년)  **2) EDA 작업**  **2.1) 자동 평점 시스템**  **2.1.1) 텍스트 정제**  - 이모지와 특수문자 제거 -> 한글, 영어, 숫자  텍스트에 이모지와 특수문자를 넣어서 같이 모델학습 할 경우 성능이 낮게 나왔기 때문에  이모지, 특수문자 정제한 텍스트 데이터를 가지고 분석하려고 정제하였음.  ->성능이 낮게 나온 이유 - 이모지 숫자가 30만개가 넘는 리뷰데이터에 비해 너무 적었고, 특수문자도 감정과는 의미없는 점과 쉼표같은 특수문자가 많았기 때문    **2.1.2) 데이터 사이즈 조절**  - 1~2점: 부정, 3점: 중립, 4~5점: 긍정  -> 1~3점: 부정, 4~5점: 긍정  - 언더샘플링 : 긍정 0.6  - 오버샘플링 : 부정 0.4  -> 긍정 데이터가 많고, 부정데이터가 적어서 비율을 맞추기위해 6:4 비율로 맞춤.  **2.2) 맞춤형 상품 추천 시스템**  **2.2.1) 데이터 슬라이싱**  - 과거 데이터를 2024년으로 줄임  -> 원본 데이터는 유지  -> 최신 데이터이므로 신뢰성 향상  **2.2.2) 데이터 통합**  - 네이버 데이터랩의 데이터를 성별로 1차  분류와 연령별로 2차 분류  -> ex) 30대 남성, 40대 남성, 20대 여성, 60대  이상 여성 등  **3) 작업 방향**  **3.1) 자동 평점 시스템**  **3.1.1) 분담**  - 김철기: 자동 평점 시스템 모델 테스트 및  성능 향상  - 박건하: 키워드 추출 및 생성 시스템 테스트  **3.1.2) 모델 설계**  3.1.2.1) klue/bert-base 학습결과  - 이진분류(sigmoid)  loss: 0.6810 - accuracy: 0.5939 - val\_loss: 0.6737 - val\_accuracy: 0.6000  ->loss가 0.68로 높았고 , 정확도, 검증정확도 0.6 성능이 낮게 나왔음.  - 회귀(mae)  loss: 1.1634 - mae: 0.8504 - val\_loss: 1.1495 - val\_mae: 0.8461  -> loss는 1.1 mae 검증 오차값이 1.1로 너무 높게나왔음.  3.1.2.2) LSTM  - 이진분류(sigmoid)    accuracy: 0.9314 - val\_loss: 0.3047 val\_accuracy: 0.9018  - 회귀(mae)  loss: 0.2914 - mae: 0.3577 - val\_loss: 0.4594 - val\_mae: 0.4461  -> 시그모이드로 LSTM학습했을때 긍부정 정확도가 전부 높게 나왔음, 검증 정확도도 0.9로 높은 성능을 보여주었음.  -> 회귀 학습결과도 낮은 오차값이 나와서 높은 성능을 보여주었음.  단, 짧은 내용의 리뷰를 예측할 땐 예측을 엄청 잘하지만 긴 문장에서 문맥을 파악할때 문맥파악은 잘 못하였음.  3.1.2.3) CNN + Bidirectional LSTM  - 이진분류(sigmoid)    loss: 0.1644 - accuracy: 0.9323 - val\_loss: 0.3006 - val\_accuracy: 0.9023  - 회귀  loss: 0.3598 - mae: 0.4391 - val\_loss: 0.4786 - val\_mae: 0.4921  -> 시그모이드로 학습했을때 loss 0.16으로 낮게나오고, 정확도 0.93, 검증정확도 0.90 높은 성능을 보여주었음.  회귀 학습결과도 mae 0.43 낮은 오차값이 나와서 높은 성능을 보여주었음.  LSTM 모델에 비해 긴 문장의 문맥 파악 능력이 뛰어나 마켓컬리의 평점 자동화 시스템에 도입 가능성이 높다.  **3.2) 맞춤형 상품 추천 시스템**  **3.2.1) 분담**  - 성원선: BERT 모델을 이용하여 Kaggle의 감정  추정 데이터 셋 테스트  - 홍석우: 샘플 데이터로 구상했던 시스템 구현,  kaggle로 데이터 맞춤 시스템 재현  - 황진웅: 마켓컬리의 밀키트 상품 정보 수집,  연령별과 성별의 음식 종류에 대한  검색량 데이터 수집(네이버 데이터랩)  **3.2.2) 진행 방향**  [사용자 입력]  └─> 날짜/시간 + 위치(수도권)  ↓  [날씨 정보 추출]  └─> 기온, 강수량, 날씨 상태 (맑음/비/눈 등)  ↓  [음식 카테고리 규칙 정의]  └─> 규칙 기반 + 머신러닝 기반 음식 분류기  (TF-IDF or BERT로 분류)  ↓  [해당 조건에 맞는 음식 추천]  └─> 상품명/설명 기반 필터링  (예: 국물 요리, 매운 음식 등)  ↓  [추천 결과 출력]  └─> 상품명, 상품 이미지, 판매 링크  **4) 금주 핵심 이슈**  [데이터 불균형]  - 평점 데이터 중 90퍼센트 이상이 긍정적인 리뷰로 데이터 불균형 발생 및 부정 데이터 학습 데이터 부족  - 해결 방안 : 부정 부분에 3점(중립)을 추가하여 라벨링과 오버샘플링  [애매한 문장]  - 맛있지만 배송이 느렸다, 포장 상태가 별로였지만 양이 많았다 등 긍정과 부정이 함께 있는 문장 분류의 모호함  - 리뷰 내용은 부정이지만 평점이 4점 이상인 경우  - 해결 방안 : LSTM과 CNN + Bidirectional과 같이 양방향 모델 사용  [맞춤형 상품 추천 시스템의 데이터 수집]  - 음식과 날씨 그리고 감정을 연결하는 데이터를 찾기가 어려움  - 해결 방안 :  1. 네이버랩에서 연령과 성별에 따른 음식 종류의 검색량 데이터 수집(2024~2025)  2. Kaggle 데이터 셋으로 학습하고 번역 모델을 이용하여 한국어-영어 연결(한국어가 입력되면 영어로 번역해서 최종 입력으로 사용) - unnatiag/korean-to-english-translator  **5) 회의 및 피드백**  [회의주제]  (1) 데이터 불균형(긍정: 90%, 부정: 10%)  (2) 모델 성능 향상을 위한 피처 선택  (3) 맞춤형 상품 추천 시스템 진행 방향  (4) 샘플링  [피드백]  (1) 데이터 불균형  -> 언더샘플링, 오버샘플링, 가중치 설정  -> 중립 데이터를 부정 데이터로 사용  (2) 피처 선택  -> EDA 작업을 통해 유의미한지 확인  -> 자모음 연속 패턴에 대한 처리 방법 확인  -> 감탄문 수에 가중치를 두는 것도 좋은 방법  (유의미한 피처에 가중치 설정)  (3) 진행 방향  -> 날씨의 과거 데이터를 통해 음식을 어떻게  추천해 주는지를 확인(데이터랩 사용)  -> 프로토타입 구현 - 수도권만 진행  (이후에 다른 지역까지 확장)  -> 상품 설명에서 상품명이 제대로 추출되는  것에 대한 정확도 확인  -> 리뷰 데이터에서 유사도를 추출할 수 있지만  리뷰 사용은 이후에 고려하기  -> 상품 정보 수집에서 컬리 팁은 제외하기  (4) 샘플링  -> 오버샘플링: 데이터량 증가  -> 언더샘플링: 데이터량 감소  -> 비율 차이가 심할 때 해당 차이만큼  오버샘플링하면 비효율적일 수 있음  (보통 2배로 설정)  **6) 인원별 세부 수행 내역**  - 김철기  : 이모지, 특수기호를 뺀 리뷰데이터 정제작업,  모델 성능 테스트, 실제 사용자 평점과, 모델의 예측평점 비교하고, 모델이 얼마나 긍부정 예측을 잘 해주는지 비교 및 자동 평점 시스템 대시보드 구현  - 박건하  : 자동 평점 시스템 모델 테스트, 키워드 자동생성 시스템 데이터 전처리 및 키워드 추출, 상품 추천 시스템 작업 진행방향 대시보드 구현  - 성원선  : Kaggle의 감정 추정 데이터 셋으로 BERT 모델과 번역 모델(unnatiag/korean-to-english-translator) 테스트  - 홍석우  : 구상하던 상품 추천 시스템을 더미 데이터를 만들어서 모델링 후 만들어 테스트 함  kaggle 데이터(Instacart, Food Demand Forecasting, Online Food Dataset)을 이용해서 다양한 수요 예측 방향성을 알아보고 테스트 함  - 황진웅  : 맞춤형 상품 추천 시스템에서 추천할 상품 목록으로 마켓컬리의 밀키트 상품 정보 수집, 기상청API로 22개 도시의 5년간 날씨 데이터 수집, 데이터랩에서 음식 종류(찌개, 국 등 - 24가지)에 대한 성별과 연령별 검색량 합계 데이터 수집(2024~2025) | **1) 자동 평점 산출 시스템**  **1.0) 작주 진행 완료사항**  - 자동 평점 시스템 구현 및 성능 향상  **1.1) 자동 평점 시스템**  - 모델 파인 튜닝  - 수집한 마켓컬리 데이터 이용해서 리뷰데이터 전처리, 리뷰데이터 분석, 시간이 허락되는 범위 내에서 모델 학습까지 수행    **1.2) 키워드 자동생성 시스템**  - keyword map 정의 및 보완  - 0/1 자동 라벨링 수행  - 학습용 csv 생성  - 라벨 결과 확인 및 키워드 점검  - SentenceTransformer 임베딩 + 분류기 학습  - 실제 리뷰에 키워드 자동 예측 적용 테스트  **2) 맞춤형 상품 추천 시스템**  **2.0) 작주 진행 완료사항**  - 마켓컬리의 밀키트 데이터 수집  - 네이버 데이터랩의 검색량 데이터 수집  **2.1) EDA**  - 날짜 기준으로 데이터를 매핑(수도권)  - 매핑한 데이터를 성별과 연령대로 구분    **2.2) 마켓컬리 상품 분류(라벨링)**  - TF-IDF 사용  - BERT 모델 사용  - 비교 후 모델 선택  **2.3) 모델 테스트**  - Kaggle의 감정 추정 데이터셋 학습  - 아마존 리뷰 데이터셋 학습  - 번역 모델 테스트  (Helsinki-NLP/opus-mt-ko-en,  unnatiag/korean-to-english-translator)  **2.4) EDA 작업방향(대시보드)**  **2.4.1) 개요**    **2.4.2) 판매분석**    **2.4.3) 리뷰분석**    **2.4.4) 날씨분석**      **2.4.5) 추천시스템** |

**2. 강사님 피드백**

| 3점은 중립이지만 부정으로 포함함'이라는 판단 기준의 명확한 근거 부족합니다. 단순히 부정 데이터가 적어서보다 구체적인 방안을 제시하면 좋을 듯 합니다.  성능 낮은 모델 BERT에 대한 분석, 해결 방안 등이 포함되면 좋을 것 같습니다.  문맥을 잘 파악한다고 평가하셨는데, Attention Layer 추가 또는 Transformer 기반 모델 활용도 고려해보세요 ㅎ  음식-날씨 매핑 데이터의 정량화도 필요할 것으로 보입니다. (신뢰도)  음식 분류기 부분도 BERT 외에도 다른 모델들을 고려하거나 멀티태스크 모델로 설계하는 것을 생각해보세요. |
| --- |