**[ 6차(7월 1주차) 프로젝트 수행 일지 ]**

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

※ 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) 자동 평점 산출 시스템**  **1.1) 자동 평점 시스템**  리뷰 길이에 따른 모델간 예측 정확도 분석  -> 리뷰길이간 분포차이  LSTM    Roberta + BiLSTM    -> LSTM의 긍정 쏠림 현상이  트랜스포머 로버타 모델로 완화되었다  리뷰 길이에 따른 예측 정확도 차이  LSTM    Roberta + BiLSTM    -> 리뷰길이 정확도도 트랜스포머모델이 안정적이게 고루 분포함  모델에 attentionlayer로 가중치 부여했을때와 부여하지 않았을때의 비교. 확인      어텐션레이어를 주었을때가 오히려 1점5점에 극단적으로 예측해버림.  어텐션레이어를 주지않는게 모델 예측이 고루분포함.      Roberta의 RMSE오차가 LSTM에비해 줄어들었음..  R2 SCORE가 0.8이상으로 준수한 성능을 보여줌.  분류모델과 회귀모델 비교 분석.      회귀모델의 분포 보정 시도                          **핵심 해석**  •모델이 리뷰 텍스트 자체보다는, 학습 데이터에 포함된 라벨의 빈도에 따라 예측 점수를 올리는 경향을 보인다.  •예: 전체 학습 데이터에 5점이 과도하게 많으면, 평점이 낮은 문장도 3~5점으로 예측.  •그래서 오차가 적어 보이지만, 사실상 리뷰 내용을 올바르게 해석하지 못하는 상태.  실제 문제  •이렇게 되면 모델이 “무조건 평균점 이상으로 점수를 주는 습관”이 생긴다.  •즉, 감정 인식이 아니라 라벨 분포 묘사에 가까워진다.  회귀모델의 범위구간 설정.    l평점 구간별 분포를 보고 범위지정,  l정수라벨 1,2,3,4,5점으로 나눔      l 리뷰가 고르게 분포하게됨  l  분류모델과 회귀모델과의 구간별 비교  .  l드라마틱한 차이는 없지만, 회귀모델쪽 구간 예측값이 조금 더 고르게 분포한다는걸 볼 수있음.  l분류모델보다 회귀모델이 출력값을 연속적으로 예측하기 때문에 자연스럽게 나타나는 현상 인 것 같다.  활용 적합성 확인    대부분 문장의 문맥을 잘 반영하여 예측해준다.  긴 리뷰 VS 짧은 리뷰 오차 비교    비슷한 경향을 보임.      같은 텍스트인데 실제평점이 다른 리뷰들 확인        실제 vs 예측의 미세한 차이  일부 리뷰는 예측평점의 표준편차가 약간 더 크거나 작다.  예측평점이 실제평점보다 분산이 약간 줄어든 리뷰도 보인다.  l즉 모델이 실제 데이터보다는 다소 중앙에 수렴하는 경향이 있다.    이 그래프는 다음 사실을 명확히 보여준다: 학습 데이터에 특정 리뷰에 1점과 5점이 함께 존재해 라벨이 분산되어 있다. 모델은 실제 분포와 유사한 경향을 보이지만, 극단 점수를 약간 보정하여 더 평균적인 예측을 낸다.  모델은 학습 데이터의 전반적인 분포 패턴을 그대로 묘사하면서도, 극단적인 케이스는 조금 완화해 예측한다.  실제평점 vs 예측평점의 패턴 유사성 리뷰 별 실제/예측 곡선 모양이 거의 비슷하다.  다만 예측평점 곡선이 전체적으로 평탄하거나 극단점이 줄어드는 경향이 보인다.  마켓컬리 리뷰데이터 적용 후 확인.        **핵심관찰(시사점)**  긴 리뷰 샘플에서도 문맥의 긍정적 표현을 효과적으로 인식하며 높은 평점을 안정적으로 부여함.  짧은 리뷰의 경우에도 단어 간 뉘앙스를 정확히 반영해 과도하게 낮거나 높은 점수를 예측하는 경향이 적음.  전반적으로 리뷰 텍스트가 내포하는 감정과 만족도를 모델이 잘 해석하여 실제 평점에 부합하는 결과를 보임.  핵심 평가  리뷰 내용과 예측평점의 일관성이 높아, 실사용에 신뢰성을 갖춘 성능을 확인할 수 있음.  예측 편향이나 극단적 쏠림 없이 전반적으로 고른 분포를 보여, 실제 구매자 만족도를 예측하는 데 유용할 것으로 기대됨.  **1.2) 키워드 자동생성 시스템**  [해시태그 모델 생성]  - 키워드/해시태그 추출 모델 적용  KeyBERT 및 SentenceTransformer 기반 핵심 키워드 추출  한글 지원 임베딩 모델 활용  리뷰별 핵심 키워드 1~3개 자동 추출  - 불용어 사전 적용  한국어 불용어 직접 수집 및 커스텀 리스트 적용  - 자동 해시태그 생성 로직 개발  해시태그 키워드 매핑 사전(csv) 설계 및 매핑 함수 구현  사전 기반 자동 태깅, 키워드 매칭 방식, 확장 표현 포함 등 반복 개선  코드로 리뷰 전처리→키워드 추출→불용어 제거→해시태그 추천까지 전체 파이프라인 구축  **2) 맞춤형 상품 추천 시스템**  **2.1) 프로토타입**  **2.1.1) 데이터 가공**  - [Food.com](http://food.com) Recipes and Interactions  - 요리과정, 레시피 설명, 식재료, 태그(레시피에  대한 짧은 설명)를 합쳐서 텍스트로 만듬  - 만들어진 텍스트를 정제함  (2g, 2~5ml, 2-4ml, ½, 2.5이런 경우를 제외한  나머지 특수기호 삭제)  - 영양성분을 칼로리, 지방, 당, 나트륨, 단백질,  포화지방, 탄수화물로 나누어서 칼럼으로 각각  저장    **2.1.2) 모델**  - 첫 번째 테스트  -> 입력 데이터: 정제된 텍스트  -> 타겟값: 칼로리 레벨(0, 1, 2)  (0: 저칼로리, 1: 중간칼로리, 2: 고칼로리)  -> 모델: BERT  - 두 번째 테스트  -> 입력 데이터: 정제된 텍스트, 영양성분(실수)  -> 타겟값: 칼로리 레벨(0, 1, 2)  (0: 저칼로리, 1: 중간칼로리, 2: 고칼로리)  -> 모델: BERT 임베딩 + 딥러닝 모델  (멀티모달 분류 모델)  **2.1.3) 결과 및 분석**  - 첫 번째 테스트  -> loss: 0.6812 - accuracy: 0.6728  -> 입력: "I'm looking for food with high calories."  -> 예측 결과: 중간칼로리  -> 분석  : 텍스트의 전처리가 제대로 안 되었음,  정보가 부족함  - 두 번째 테스트  -> loss: 0.1323 - accuracy: 0.9488      -> 입력: "This dish is rich in cream and cheese",  [10, 8, 600, 20, 5, 45]  [지방, 당, 나트륨, 단백질, 포화지방, 탄수화물]  -> 예측 결과: 고칼로리  -> 분석  : 타겟값인 칼로리와 유의미한 상관관계가 있는  영양성분을 피처로 추가해서 사용하니 성능이  향상되었음, 저칼로리를 분류 못하는 문제는  데이터의 수가 적어서 생기는 걸로 보임    **2.2) 문장 생성**  **2.2.1) 목표**  **- 챗봇 로직**  -> 1~2번의 대화 후에 음식을 추천  -> 사용자의 응답에서 특징을 추출하여 유사한 제품  5가지를 추천  -> 사용자의 요구사항을 담겨있는 문장 데이터 필요  -> 음식 종류 카테고리, 태그, 감정, 식재료로 문장을  생성(1~4개 조합)  -> 레시피에 따라 설정된 조합으로 문장 60,000개 생성  -> 중복 제거한 전체에서 10,000개 생성(다양한 표현)  -> 문장 생성 비율: 1개 조합 - 30%, 2개 조합 - 40%,  3개 조합 - 20%, 4개 조합 - 10%  **2.2.2) 데이터 가공**  -> 카테고리, 태그, 식재료, 감정이라는 칼럼을 만들고  음식 종류 카테고리를 기준으로 태그, 식재료,  감정을 모은 후에 중복 제거하여 값을 대입  -> 카테고리, 태그, 식재료, 감정 데이터를 중복  제거하여 고유 데이터를 수집  (문장의 다양한 표현을 위해 수집)  **3) 금주 핵심 이슈**  [챗봇]  - 로직 설계 필요  - 메인 기능으로 빠른 작업 필요  **4) 회의 및 피드백**  [회의주제]  (1) 6월 30일 : 진행 상황  1) 자동 평점 시스템  -> 최적의 모델 선정 - 결과 비교  2) 해시 태그  -> 사전을 만들어서 진행을 하였는데 사전 없이  자동으로 할 수 없는지?  -> 비슷한 태그들만 출력되는 문제 발생  3) 맞춤형 추천 시스템  -> 칼로리 레벨을 타겟값으로 두고 진행  (2) 7월 1일 : 진행 상황  1) 자동 평점 시스템  -> 회귀 모델의 결과로 나온 예측 평점을 소수점  첫째자리까지 반올림해서 사용  (3) 7월 2일 : 진행 상황  1) 맞춤형 추천 시스템  -> 추천 시스템 모델의 학습 데이터로 사용할 문장  생성 방법  -> 문장 생성을 위한 요소 선택  (태그, 식재료, 감정)  [피드백]  (1) 6월 30일  1) 자동 평점 시스템  -> 표면적인 부분은 괜찮음  -> 트랜스포머가 분류 성능이 좋음  -> 회귀 모델의 성능이 왜 낮게 나오는지 고찰과  검증이 필요  -> 분류 자체는 기본적으로 argmax로 정수로  나오기 때문에 성능이 높지만, 회귀로 하면  실수로 나오기 때문에 성능이 낮아질 수 밖에  없음  -> 리뷰처럼 한 쪽으로 편향되기 쉬운 데이터이므로  처음에 분류를 진행하고 그 이후에 회귀로  미세조정하면서 진행하면 됨  (실무에서도 똑같이 진행)  2) 해시 태그  -> 해시 태그가 있는 사이트에서 크롤링해서 내용과  태그 데이터도 함께 가져와서 진행  -> 이미 학습된 모델(외부 모델)에서 해시 태그를  추출하는 방법이 있음  -> 해시 태그는 멀티 라벨링 작업  (어떤 부분으로 나누어 줄 지가 필요)  -> 같은 태그만 나오는 이유  : 리뷰 데이터와 해시 태그의 데이터가 적어서  문제가 생김(데이터 부족)  -> 실제 리뷰에서 어떤 태그가 붙는지 조사해볼  필요가 있음  -> 실제 데이터를 전처리해서 진행  3) 맞춤형 추천 시스템  -> 현재는 피처로 사용된 데이터 종류가 너무  많아서 저칼로리 데이터를 제대로 분류 못함  -> 챗봇의 입력 데이터와 출력 데이터에 대한  로직을 정하면 좋음  (챗봇과의 대화가 어떻게 진행될 것인지 정해보면  좋음)  -> 실제 입력 받을 데이터 형식으로 학습시키면  좋을 것 같다.(문장 생성이 필요)  (2) 7월 1일  1) 자동 평점 시스템  -> 소수점 첫째자리까지 내는 것이 트렌드라는 점을  근거하는 내용이 부족함  -> 실제 학습으로 사용했던 데이터에 맞게끔  조정해보면 좋음  (3) 7월 2일  1) 맞춤형 추천 시스템  -> 챗봇을 통해 나온 문장을 무엇과 매핑을 할 지와  같은 시나리오를 설정하는 것이 좋음  -> 직접 만든 데이터로 진행하면 좋음  -> 번역기는 API랑 NLP 모델에서 찾으면 좋음  (어떻게 번역되는지도 체크 필수)  **5) 인원별 세부 수행 내역**  - 김철기  모델 간 성능 비교 분석   * LSTM 분류모델, RoBERTa + BiLSTM 분류모델, 회귀모델을 학습·평가 * 각 모델의 혼동행렬과 예측 분포 그래프를 통해 정확도, 편향, 예측 경향 차이를 비교. * 리뷰 길이에 따른 모델 간 예측 정확도 분석 * Attention Layer 가중치 부여 유무 비교 * 분류모델과 회귀모델 비교 분석   평점시스템에 적합한 모델 선택 도출  회귀모델 분포 보정 시도  회귀모델 범위 구간 설정  마켓컬리 리뷰데이터에 적용후 검증.  - 박건하  : 리뷰 데이터의 해시태그 자동 생성 시스템 구축을 목표로 키워드 추출 모델(KeyBERT, 임베딩), 불용어 처리, 해시태그 매핑 로직을 개발하였으며 반복적 피드백을 통해 성능 및 품질 개선을 추진  - 성원선  : 대량 데이터셋 전처리 후 googletranslator를 이용하여 번역작업 (많은시간이 소요됨)  번역모델 처리 진행과 동시에 챗봇 구현관련 작업을 위한 코드 (streamlit or UIzard 등 사용예정) 작업 진행중  더불어 컬리 식품 페이지 내 평점시스템등 시각화 구상중에 있음  - 황진웅  : 문장 생성을 위한 데이터를 가공하였음  가공한 데이터를 바탕으로 60000개의 문장 데이터를 생성 | **1) 자동 평점 산출 시스템**  **1.0) 작주 진행 완료사항**  - 자동 평점 시스템 구현 및 성능 향상  - 부정리뷰 데이터 확보 및 데이터 비율 조정  - 모델의 정확도 저하 원인 파악을 위한 EDA  -> 키워드별로 쪼개서 평점 1,2,3,4,5점의 실제 긍/부정/중립 리뷰 키워드의 비율 확인  -> 평점 3점에 중립적인 표현보다는, 긍정,부정표현과 섞여있음을 확인.  **-** 긍정 부정 감정을 판단내리지못하는 중립적인 표현들 전처리. 전체 평점키워드 파악, 감정어, 강조어, 명사로 감정에 영향을 끼치는 지 확인.  ->Keyword 기반 감정 점수 한계점: 문맥을 반영하지 못함을 확인.  평점기반 라벨링을 기준으로 삼아서 재분석  -> 모델 학습하면서 EDA작업 진행함  LSTM보다는 트랜스포머모델(Roberta)+BiLSTM 혼합한 모델이 성능이 전체적으로 좋음을 확인  [자동평점 시스템]  대시보드 + UI 디자인 설정  [해시태그 자동생성 시스템]  - EDA 및 초기구축  - 구조 전환 및 기능 구현 테스트  - 사전 기반 자동 태깅 및 성능 개선 반복  **1.1) 키워드 자동생성 시스템**  - SentenceTransformer vs KoSBERT 등 한국어 특화 임베딩 비교  - 평가 메트릭(코사인 유사도, 매칭률 등) 확인, 추후 수작업 평가/자동평가 도입 필요  - 해시태그 자동생성 시스템 고도화를 위해 카테고리별 해시태그 분류 체계 수립  - 라벨링된 리뷰-해시태그 데이터셋 (200~300개) 구축 및 지도학습 기반 멀티라벨 분류 모델 적용 준비  - 해시태그 임베딩 품질 및 편향·중복 여부 EDA  - 최종적으로 성능 평가 체계 도입 방안 검토  **2) 맞춤형 추천 시스템**  **2.0) 작주 진행완료 사항**  - 프로토타입 테스트 완료  - 60,000개의 영어 문장 생성 완료  **2.1) 문장 생성**  - 전체 요소의 고유 데이터로 문장 10,000개  생성  - 생성한 문장을 한국어로 번역해서 확인  - 챗봇을 구현하기 위한 한국어 문장 생성  **2.1.1) 레시피 정보 번역**  영어로 되어있는 20만여개의 레시피 정보 데이터 번역  음식명 -> 마켓컬리와 맵핑  **2.2.2) 챗봇 로직 설계**  키워드 기반 BIO 태깅 로직 시도해봄.  챗봇 응답 패턴 시나리오 구축  정규식 및 유사어 매칭 룰 최적화  **2.2) 테스트**  - 영어 문장 생성 결과 확인  - 영어 문장 데이터를 이용하여 추천 모델 학습  진행  (X: 영어 문장, y: 카테고리/태그/식재료/감정)  - 챗봇 동작  **2.3) 데이터 수집**  - 마켓컬리의 밀키트와 연결하는 방법  **2.4) 모델 선정**  - BERT, KoBERT  - 코사인 유사도  - 앙상블(XGBoost, Light GBM, CatBoost)  - 번역모델  facebook/nllb-200-distilled-600M |

**2. 강사님 피드백**

| 비교 부분에서는 정확한 수치데이터를 토대로 비교를 나타내주시면 좋을 것 같습니다. 이후 다른 부분과 비교할 때도 활용이 가능합니다.  키워드 자동생성 시스템 부분은 생략된게 많네요. 이전에 말씀드린 것처럼 평점 시스템 참고해서 자세히 진행하는 것이 좋을 것 같습니다.  추천시스템 또한 생성 문장의 예시와 방식 및 데이터 가공 부분의 시각화 자료 등 진행 사항을 상세하게 나타내주는 것이 좋습니다.  실패한 시도들도 왜 시도했는지, 왜 실패한건지, 무엇을 배웠는지 등을 정리해두시면 보완하여 사용하거나 이후 프로젝트의 근거자료나 포트폴리오에서도 사용가능합니다.  뿐만 아니라 여러분들이 여러 시도 끝에 잘 해결한 부분들도 상세하게 기록해두면 이후에 사용하기 좋습니다. |
| --- |