**[ 5차(6월 4주차) 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) 자동 평점 산출 시스템**  **1.1) 자동 평점 시스템**  작주에 긍정에 편향되었던 리뷰데이터를 해결하기 위해 오버샘플링을 돌려서 어느정도 해결을 보았었지만 여전히 특정 리뷰데이터들은 예측이 빗나감.  -> 샘플링보다는 실제 데이터의 부정 리뷰데이터를 가져오는게 더 모델성능 정확도가 높게 나올것이라는 피드백을 받음 그래서 부정리뷰데이터를 더 확보함.  -> 카카오맵의 실제 부정리뷰데이터 7만개를 확보하였고, 수집한 11번가, 옥션의 데이터와  카카오맵의 7만개 부정데이터를 합침.  -> 그 결과 긍정쪽에 편향되었었던 리뷰데이터가 부정쪽에도 어느정도 비율을 맞춤.    평점별 리뷰수 확인    평점별 리뷰길이 분포 확인    이상치부분은 제거해줌    리뷰데이터터-> 키워드로 접근해봄  감정어 상위 30개 분포도    평점별 감정어 상위10개 분포도    설명 :  전체 키워드 감정어 상위30개의 그래프를 보면  좋아요,잘,맛있어요,감사합니다, 와 같은 긍정표현 분포가 확실히 높게 나옴을 알수있었다.  -> 긍정 데이터 많음  평점 1~2점에는 없고, 했는데, 없어요, 없네요, 와 같은 부정 표현키워드가 많이 나타났다.  평점 4~5점에는 좋아요, 맛있어요, 감사합니다, 빠르고, 저렴해게 와 같은 긍정키워드가 많이 나타났다.  반면 평점 3점에는 좋아요, 잘, 했는데, 재구매, 않고, 와 같은 긍정,부정 표현이 섞어있고 재구매의 단어가 많이 나타났다,  ->중립적인 표현의 단어가 거의 없었다.  강조어 빈도분포    평점별 감정어 상위10개 분포도    설명  강조어  너무,많이,다,아직,없이,엄청,빨리, 그럭저럭, 와 같이 감정판단에 영향가는 키워드와 영향 안가는 키워드들이 섞어있었다.  평점 1~5점 모두에 너무,많이 라는 단어가 다 높게 분포되어있었고,  1~2점에는 없이 라는 단어가 조금 분포되어있는 반면에  4~5점에는 아직, 그래도 라는 단어가 많이 있었다.  3점에는 모든 단어가 골고루 다 분포되어있다.  명사      -> 긍부정 감정 평가하기에는 전혀 관련없는 단어들이다.  평점기반의 라벨로 학습하는것보다 실제 텍스트 기반라벨을 생성해서 학습해보는것이 더 근거있지않을까 해서,  형태소분석기로 리뷰데이터를 키워드로 쪼개어  긍부정표현의 키워드들을 분석하였고, 군산대학교 연구실에서 만들어진 감정사전과 맵핑해서 감정점수를 합산하는 방식으로 강한긍정, 약한긍정 중립 약한 부정, 강한 부정 이런식으로 라벨을 생성해 학습하려고 했지만, 키워드의 감정점수가 오히려 민감하게 반응해서 리뷰내용의 문맥을 파악하지 못하였다.  **핵심 한계: 문맥을 반영하지 못하는 Keyword 기반 감정 점수**  키워드는 문장 내 위치나 연결어의 뉘앙스를 고려하지 못함. "배송은 빨랐지만 제품은 실망" 같은 문장에선 후반 정보가 핵심인데 키워드 감정 합산은 이를 구분 못함. 감정 사전도 단어별 극성은 있되, 강조 표현/전환어/아이러니 등을 처리하지 못함  그래서 다시 정답지가 있는 평점라벨 기반으로 분석 및 예측하는 방법으로 전환하였다.  LSTM 평점기반 긍부정중립 3라벨 학습결과    혼동행렬    -> 긍부정 90%이상 성능을보였고, 중립은 55%로 성능이 낮았다, 혼동행렬도 긍부정에 잘 예측한 반면 중립은 분산되어있음을 확인할수있다.  리뷰길이분포확인      부정, 중립, 긍정 순으로 리뷰길이가 길게나왔고,  리뷰길이 구간별 예측정확도는 0~50에 약간떨어짐, 100~150구간 상승, 151~부터는 정확도가 상승하였다  감정예측 정/오답 분포    -> 긍정 부정에는 실제감정과 빗나간게 거의 없고, 중립은 실제감정과 많이 빗나가게 예측된걸 알수있다.    예측정확도도 중립이 현저하게 낮았다.  LSTM회귀모델 긍/부정/중립 결과확인        -> 중립쪽의 정확도가 살짝 올라갔고, 나머지는 비슷하다.  LSTM 분류모델 -> 평점1~5점 라벨 기준결과      평점 4점 5점에서 실제값과 예측값이 많이 달랐다.  실제 5점 비율이 매우 높지만 예측에서는 4점으로 과다하게 쏠림  → 모델이 최고 평점(5점)을 과소예측하고 있음  2점, 3점은 전체적으로 과소예측 예측에서 2점, 3점은 실제에 비해 낮게 나옴  실제 평점의 예측평점 분포    실제 평점 3점의 예측평점 분포    -> 평점 1~2점은 예측평점이 실제평점과 65%이상으로 높게나왔고,  4점도 65%, 5점은 평점4점이 46% 5점이 50%로 거의 반반비울로 나옴,  3점은 예측평점3점이 36%로 낮게나옴,  정밀도 기반 예측VS실제 히트맵  ->회귀모델  평점 구간별 예측값.    혼동행렬    -> 실제 3점구간쪽은 예측값이 낮음을 알수있음  리뷰 길이구간별 RMAE/MAE 오차그래프    -> 리뷰길이가 0~100구간에서 점점 줄어들다가  150부터 조금 상승하는가싶더니 201구간부터 다시 줄어듬    -> 3점쪽에 확실히 예측값이 떨어짐  - LSTM + Transformer모델로 혼합 학습(Roberta + BiLSTM) -> 긍부정중립라벨    LSTM모델과 비슷하게나왔지만, 혼동행렬을 보면  성능은 확실히 올라감.    잘 예측함  감정별 정/오답 분포    LSTM모델에비해 오차가 줄었음 특히 중립쪽이 개선되었음.  감정별 정확도    정확도가 LSTM에비해 전체적으로 많이 올라갔음    리뷰길이 분포는 LSTM과 비슷함    리뷰 구간별 길이 예측 정확도도  LSTM에비해서 균형있게 분포된 것 같다.  **1.2) 키워드 자동생성 시스템**  [해시태그 모델 생성]  - 기존 KoGPT 기반 Few-shot 방식 해시태그 자동 생성 모델 성능 미흡으로 사전정의된 해시태그 사전을 이용한 유사도 추천 시스템 구조 전환  - SentenceTransformer를 활용해 리뷰와 해시태그 키워드 간 임베딩 비교 및 cosine similarity 계산  - 해시태그 벡터 정규화 적용으로 유사도 계산 안정성 향상  - 임계값 조정 기능 추가로 추천 민감도 튜닝 테스트 및 리뷰 감정 분석과 맛 관련 태그 필터링 적용 여부 확인 중  **2) 맞춤형 상품 추천 시스템**  **2.1) Kaggle 데이터셋**  **2.1.1) 데이터**  - [Food.com](http://food.com) Recipes and Interactions  - [Instacart Market Basket Analysis (보조)](https://www.kaggle.com/datasets/psparks/instacart-market-basket-analysis)    **2.1.2) 기능**  - 사용자의 취향을 반영하여 음식 추천  (식재료, 칼로리, 조리 과정 정도)  - 테마 기반 음식 추천  (파티, 캠핑, 비건, kid-friendly, 나라별 음식 등)  - 사용자의 현재 상태를 반영하여 음식 추천  (두통, 복통, 감기, 화남 등)  **2.1.3) EDA**  - RAW\_recipes.csv  - **칼럼 설명**  - name : 레시피명(음식명과 같은 경우도 존재)  - id : 레시피ID  - tags : 해당 레시피의 음식 특징  (kid-friendly, serve-hot, serve-cold, party 등)  - nutrition : 영양 표시  (칼로리, 지방, 당, 나트륨, 단백질, 포화지방, 탄수화물)  - n\_steps : 조리 단계 수  - description : 레시피 설명  - ingredients : 레시피에 필요한 식재료  - **활용 방향**  - tags : 552개, 계절, 따뜻한 음식, 찬 음식, 비건, 캠핑,  파티, 질병 등 일상생활에서 찾을만한 주제가 있다.    -> preparation과 time-to-make와 course가 많음    -> 질병에 관한 데이터가 없거나 극소수라서  사용자의 건강 상태는 사용할 수 없음  -> 감정 관련 데이터가 충분히 있으므로 사용자의  감정에 따라 음식을 추천해줄 수 있음  - nutrition : 영양 표시가 있어서 다이어트와 식단  추천도 해볼 수 있다.      - n\_steps : 조리 단계가 수치로 되어있어서  음식(레시피)을 추천할 때 요리 복잡도도 적용 가능      **3) 금주 핵심 이슈**  [데이터검증]  - 외부 데이터와 연결하려면 해당 데이터가 유효한 데이터인지를 검증이 필요하다.  (LampCook와 CNN의 조사 자료 등)  **4) 회의 및 피드백**  [회의주제]  (1) 26일 : 진행 상황  1) 자동 평점 시스템  -> 긍정, 부정, 중립 예측 정확도 문제  2) 맞춤형 추천 시스템  -> [Food.com](http://food.com) Recipes and Interactions EDA 결과  -> [Instacart Market Basket Analysis](https://www.kaggle.com/datasets/psparks/instacart-market-basket-analysis) EDA 결과  [피드백]  (1) 26일  1) 자동 평점 시스템  -> 텍스트가 아닌 평점 기반으로 진행  -> 회귀 + 규칙으로 평점 부여  -> 트랜스포머나 BERT를 사용하면 문맥을 파악할 수  있다.  -> 데이터를 전처리하는 것보다는 모델을 조정하는  것이 중요하다.  -> 텍스트를 정제하고 모델을 튜닝해서 학습시키면  해결할 수 있다.  2) 맞춤형 추천 시스템  -> 공통: 데이터에 대한 검증이 필요하다.  -> 식재료에 대한 데이터셋 검증이  필요하다.(LampCook, CNN 자료)  -> 칼로리와 영양 성분으로 음식을 추천해줄 수  있다.  -> 조리 난이도는 일반화가 되어야 한다.  (회는 요리 단계가 적지만 요리하기가 힘들다.)  -> 데이터를 살펴본 느낌이 강하다.  -> 특정 목적으로 요리한다고 생각하면 tag를  사용할만하다  -> 칼로리, 영양 표시, 태그, 레시피 설명으로  감정과 식단 관리 부분은 괜찮아 보인다.  -> 입력을 사용자가 직접 옵션을 선택하는 것으로  한다면 제한적일수도 있다.  -> 사용자가 문장을 입력해서 받으면 좋을 것 같다.  나중에 챗봇으로 사용 가능하다.  **5) 인원별 세부 수행 내역**  - 김철기  [키워드 기반으로 EDA 진행]  1. 형태소분석기로 쪼개서 , 감정어, 강조어, 명사 구간별 시각화 비교 분석.  -> 평점별로 실제 긍부정 키워드가 얼마나 있는지 분석하기 위해 진행함.  2. 결론도출 -> 키워드 기반으로 실제 만들어진 감성사전과도 맵핑해서 분석도 해봄 ,  실제 키워드는 리뷰의 전체문장의 문맥을 파악하지 못하는 문제점이 있음.  3. 평점기반으로 모델학습하며 EDA 진행  3.1 [LSTM 모델 학습 진행]  -> 긍부정중립 3라벨, 평점 5라벨 성능 비교분석  -> 다중분류와 회귀모델 성능비교분석  3.3 [Robert+BiLSTM 모델 학습 진행]  -> 긍부정/중립/ 3라벨 학습진행  비교 분석 시각화  - 박건하  [해시태그/키워드 자동 생성 시스템 구축]  : KoGPT 기반 해시태그 자동 생성 모델 프롬프트 최적화 작업 중 성능 저하로 작업방향 변경. 해시태그 사전을 만들어 사전에 정의된 후보군과 리뷰 문장 간 의미적 유사도를 비교하는 방식으로 시스템 개선. 이후 SentenceTransformer 기반 임베딩, 리뷰 감정 분석, 맛 관련 태그 필터링 등 적용 여부 테스트.  - 성원선  : 작주 데이터셋 EDA 작업 후 감정분석 Kluebert & Robert를 이용해 진행 해보았으나 중립 분석은 여전히 미비하여 성능을 높혀보려고 3class를 이용하여 작업하여 시각화 대시보드 준비중 / 감정 텍스트 & 평점 점수 합산을 도입해보고자 테스트중  - 홍석우  : kaggle 데이터 instacart market을 EDA 작업하여 레시피 기반, 감정 기반 추천에 기능 추가하는 방향을 모색함  - 황진웅  : kaggle 데이터 [Food.com](http://food.com) Recipes and Interactions을 EDA 작업 완료 | **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  [해시태그 자동생성 시스템]  - EDA 및 초기구축  - 구조 전환 및 기능 구현 테스트  **1.1) 자동 평점 시스템**  평점 1~5점 다중분류로 Roberta+ BiLSTM학습,  LSTM의 평점 다중분류와 성능비교 분석  회귀모델로도 학습해보고 분류랑 성능 비교분석해봄  중립을 뺀 이진분류 모델로도 학습해보고  성능 비교분석해봄  성능이 더 잘나오는쪽으로 마켓컬리 데이터와 연동.  **1.2) 키워드 자동생성 시스템**  [키워드 분류모델 생성 : 일정 지연으로 이월]  - KoBERT + 다중 라벨 활용  - 키워드 정의 및 확장  - 리뷰별 자동 라벨링 (0/1)  - 학습용 데이터 생성  - 라벨링 결과 검토  - 문장 임베딩 및 분류기 학습  - 신규 리뷰 예측 적용  **2) 맞춤형 추천 시스템**  **2.0) 작주 진행완료 사항**  - [Food.com](http://food.com) Recipes and Interactions 데이터  EDA 작업  - instacart 데이터셋 EDA 작업  **2.1) EDA**  - 칼로리로 구현한 추천 시스템의 결과 분석  - 영양 표시와 태그 그리고 레시피 설명을  피처로 추가한 작업에 대한 결과 분석  - 데이터에 대한 확실한 검증을 통해 근거있는 데이터만 사용  **2.2) 테스트**  - 칼로리로 추천 시스템 구현  - 영양 표시와 태그 그리고 레시피 설명을  피처로 추가해서 작업  **2.3) 데이터 수집**  - 음식 선호도 관련 데이터  - 마켓컬리의 밀키트와 연결하는 방법  **2.4) 모델 선정**  - 나이브 베이즈(데이터가 독립적이어야 함)  - SVC(sigmoid 커널)  - 맨하탄 거리(유사도)  - 앙상블(XGBoost, Light GBM, CatBoost) |

**2. 강사님 피드백**

| 예측 실패 원인이 명시되었으면 좋겠습니다. 가정해보는 것도 좋으니 고민해봅시다.  모델 비교 결과 요약 등이 포함되어 있으면 좋을 것 같습니다. 그래야 같은 결과를 보고 있는지를 인지할 수 있을 것 같습니다.  EDA를 포함해 결과를 그려놓고 진행을 해주시는 것이 좋을 것 같습니다. 모델 사용하기전에 데이터에 대한 분석과 목적과 목표를 갖고 학습할 데이터와 추출할 데이터를 명확히 해줍시다.  회의 때 얘기한 것을 더 중점적으로 다음주에 모델들을 사용해서 결과 확인이 되었으면 좋을 것 같네요 ㅎ 주말에도 막히시면 언제든지 질문해주세요. |
| --- |