**[ 4차(6월 3주차) 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.    - LSTM + Transformer모델로 혼합 학습(Robert + BiLSTM)  -> 긍정, 부정에는 정확도가 높게 나왔지만 중립에는 정확도 성능이 낮게 나와서 원인 파악하기위해 EDA작업 진행.  -> EDA작업  실제 평점 리뷰데이터에 긍정/부정/중립 비율을 파악하기 위해 리뷰 데이터를 키워드로 분리해서  평점별로 키워드 빈도수를 확인함      평점 1점과 2점 , 4점과 5점에는 어느정도 긍부정 데이터가 평점에 맞게 비율이 분포되어있었음.  하지만 3점에는 긍정, 부정표현의 키워드가 섞여있음을 확인할수있었음.    ex) 긍정표현 : ‘좋다’ ‘맛있다’ ,부정표현 : ‘없다’, ‘작다’  평점 3점의 데이터를 심층 분석하기 위해 표준국어대사전에있는 긍정 부정 키워드를 불러와서 Word2vec를 이용해서 비지도 학습을 시켜봄    임계값을 안주고 했을때 평점 4점5점에 긍정표현 90%이상이 나왔고, 평점 1점 2점은 부정은 80%정도 나옴. 평점3점은 긍정66%, 부정 33%,  그리고 모든 평점에 중립이 1% 미만대로 나와버림.    모든 평점에 중립적인 표현이 1% 미만대로 나와버린 이유와, 특히 평점 3점대에 중립적인 표현이 거의없고 긍정 부정 비율이 높게 나타난게  임계값을 안줘서 그런거 아닌가 싶어서 임계값을 많이도 줘보고 낮게도 줘봄.    -> 임계값을 많이 주었을때의 그래프인데 높아질수록 평점 3점의 불일치율은 낮아지고, 1,2점 4,5점의 불일치율은 올라감.  -> 모델이 긍부정 표현까지 중립으로 해석해버림    임계값이 내려갈수록 평점 1,2점 4,5점은 불일치율이 내려갔고, 평점3점에는 불일치율이 높게나왔음.  -> 임계값이 낮을수록 모델이 긍정, 부정 표현은 잘 분류한다는것을 알수있었고, 긍정/부정이 섞인 3점 중립은 확실히 구분 못하였음.  임계값을 0.01로 설정하고 돌린결과, 중립 비율이 조금 올라갔고, 평점 1,2점 불일치 비율도 조금 올라갔음.    평점 3점의 긍/부정 중립의 비율은 여전히 긍정 부정표현의 비율이 높았고, 중립 표현비율은 낮았음.    -> 평점의 긍부정 중립의 비율은 어느정도 알수있었지만, EDA작업중에 놓친 부분이 있었다. 그래서 비지도 학습을 했을때 중립의 비율이 낮게 나올수밖에 없었다는 피드백을 받았고, 실제 리뷰데이터의 중립적인 표현이 어느정도 있었는지 좀 더 자세한 EDA작업과 데이터 심층 분석이 필요하다는 피드백을 받았고, 긍/부정/중립의 키워드에 비중을 많이차지하면서, 표현에 영향을 안주는 것들은 전처리 작업이 더 필요하다는 피드백도 받았다. ex) 하다, 먹다  실제 중립적인 표현에 긍정부정표현자체를 알 수 없는 데이터도 있었고 이 부분도 전처리가 필요할 것 같다.    **1.2) 키워드 자동생성 시스템**  [해시태그 모델 생성]  - KoGPT 기반 해시태그 자동 생성 테스트  - Few shot leariing 모델학습  - 해시태그 웹 앱 프로토타입 실험  - 리뷰 클러스터링 분석  - 해시태그 품질 향상 테스트    **2) 맞춤형 상품 추천 시스템**  **2.1) 마켓컬리 + 데이터랩**  **2.1.1) 데이터**  - 날짜 기준으로 데이터를 매핑(22개의 도시)  - 매핑한 데이터를 성별과 연령대로 구분  - 학습데이터 (성별/연령대/지역/기온/습도/하늘상태/불쾌지수)  - 타겟값: 음식 종류  - 성별, 연령대, 지역을 랜덤으로 선택하여 학습  데이터 생성    **2.1.2) 마켓컬리 상품 분류(라벨링)**  - TF-IDF 사용  - KoBERT 모델 사용  - 비교 후 모델 선택  **2.1.3) 모델 테스트**  - Kaggle의 감정 추정 데이터셋 학습  - 번역 모델 테스트  (Helsinki-NLP/opus-mt-ko-en,  unnatiag/korean-to-english-translator)  **2.1.4) 한계**  - 데이터랩의 수치가 절대값이 아니라 상대적인  값이라서 카테고리간의 비교 불가능  **2.2) Kaggle 데이터셋**  **2.2.1) 데이터**  - [Food.com](http://food.com) Recipes and Interactions  - [Instacart Market Basket Analysis (보조)](https://www.kaggle.com/datasets/psparks/instacart-market-basket-analysis)    **2.2.2) 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)  **2.2.3) EDA**  - RAW\_recipes.csv  - 필요없는 칼럼: minutes, contributor\_id, submitted,  steps, n\_ingredients  - **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.  - nutrition : 영양 표시가 있어서 다이어트와 식단  추천도 해볼 수 있다.  - n\_steps : 조리 단계가 수치로 되어있어서  음식(레시피)을 추천할 때 요리 복잡도도 적용할 수  있다.  **3) 멘토링 피드백**  **임호태 멘토**  [추천 시스템의 한계 지적 및  기업 활용도 중심 인사이트 강화 강조]  - 추천 시스템 : 나이, 성별, 날씨 외 지역, 감정, 구매이력 등 다양한 변수로 개인화 강화 필요  - 프로젝트 보완 : 리뷰 인사이트 확장 및 제품별 6각형 그래프 등 시각화를 통해 기업 활용도 중심 설계 제안  - 포트폴리오/면접 : 발표자료에 기여도와 성과 명확히 정리와 면접 대비 경험 기반 스토리 숙지  **김보경 멘토**  [모델링 데이터 처리 정확성 및  포트폴리오 구성과 논리적 흐름 강조]  - 추천 시스템 : 감정, 음식 선호 등 사용자 정보를 추가해 추천 정확도를 높이고 인기 기반 랜덤 추천도 고려  - 데이터 처리 : 오버샘플링 시 단순 복제 대신 문장 다양화로 과적합 방지 필요  - 포트폴리오/면접 : 문제정의부터 활용까지 논리 흐름 강조 및 짧고 명확한 결과 중심 구성  **4) 금주 핵심 이슈**  [EDA]  - 데이터 분석의 설득력과 분석 결과의 신뢰도를 높이기 위해 인사이트 도출의 기반이 되는 탐색적 데이터 분석을 보다 체계적이고 정교하게 수행 필요 (EDA 과정에서 변수 간 상관관계와 데이터 분포에 대한 심층적 검토)  **5) 회의 및 피드백**  [회의주제]  (1) 17일 : 진행 상황  1) 자동 평점 시스템  2) 맞춤형 추천 시스템  -> 성별, 나이, 지역, 기온, 습도, 하늘상태, 불쾌지수로  학습 데이터로 사용  -> 데이터랩의 검색량 데이터 기반으로 타겟값 매핑  (2) 18일 : 진행 상황  1) 자동 평점 시스템  -> LSTM, CNN + BiLSTM, BERT 모델 학습 결과  2) 맞춤형 추천 시스템  -> 카카오맵 데이터  -> kaggle 레시피 데이터  (3) 20일 : 진행 상황  1) 자동 평점 시스템  -> 평점 데이터들을 EDA한 결과  -> 중립 데이터를 처리하는 방법  [피드백]  (1) 17일  1) 자동 평점 시스템  2) 맞춤형 추천 시스템  -> 데이터랩 : 단위량 고려 -> 절대값인지?  -> 데이터 분석을 더 해야한다.(EDA)  -> 특정 features에 과적합이 생긴 것 같다.  -> 학습 데이터 체크  -> 개인화 문제: 똑같은 것을 추천해줄 수 있다.  -> kaggle의 거래 데이터를 사용  -> 마켓컬리 내에서 해외 음식을 추천하는  프로모션 고려  (2) 18일  1) 자동 평점 시스템  -> 긍정, 부정, 중립 모든 데이터를 더 자세히  분석해봐야 한다.  -> 중립 데이터가 왜 분석이 제대로 안 되는지  분해해서 찾아보기 -> EDA를 진행  -> 뽑아낼 수 있는 정보를 찾아보자 -> 리뷰 말고도  외부적인 요소를 가져와서 확인해보기  2) 맞춤형 추천 시스템  - 카카오맵 데이터  -> 리뷰 수는 해당 날짜에 어떤 음식을 많이  원하는지를 알 수 있는 용도로 사용할 수 있다.  -> 특정 지역에 어떤 음식의 음식점이 많다.  프로모션으로 활용하면 이용할 가치가 있다.  - kaggle 레시피 데이터  -> 어떤 모델을 사용할지를 명확하게 정하면 좋다.  -> 굳이 학습이 필요할까? 조건문으로 해결되지  않을까?  -> 같은 음식만 추천할 우려가 있을 것이다. 어떻게  해결?  -> 정답이 있고 예상이나 예측할 수 있는 것을  AI를 이용해서 예측하는 것이 좋다.  (3) 20일  1) 자동 평점 시스템  -> 데이터 출처를 명확히 알고 있어야 한다.  -> 표준어 사전에서 중립 단어도 함께 가져와야  한다.  -> 키워드를 추출했으면 각 점수에서 공통적으로  나오는 키워드를 제거하면서 EDA 진행해보기  -> 중립 키워드와 중립 리뷰 내용을 보고 해석할 수  있는지 확인  -> 회귀로 진행하고 난 후에 해석과 분석해서 평점  부여하는 범위를 정하면 좋겠다.  -> EDA 진행과 규칙 추가  **6) 인원별 세부 수행 내역**  - 김철기  [Robert+BiLSTM 모델 학습 진행]  : 평점 기반의 감정 라벨링 생성 후,  이진 분류 및 회귀 실험  - EDA 작업 진행  : 중립표현의 성능 미흡 판단 -> 1~5점 평점 구간별로 리뷰데이터의 긍정,부정,중립 키워드 확인, 리뷰 실제 평점 vs 텍스트 키워드 비교  평점별 키워드 빈도수 시각화  Word2Vec 비지도학습 후 결과 비교  평점과 텍스트 예측값 간 불일치 체크,  임계값 최적화와 중립 필터링 조건 검증을 위해 임계값구간 설정해서 0.005~ 0.5까지 구간대비 긍/부정/중립 예측값 비율 비교,  - 박건하  [해시태그/키워드 자동 생성 시스템 구축]  : Okt 형태소 분석과 TF-IDF 기반 KMeans 클러스터링을 통해 리뷰를 주제별로 분류하고 대표 문장 추출, KoGPT 기반의 few-shot 프롬프트를 설계해 리뷰 문장에서 자연스러운 해시태그를 생성, 결과 품질을 높이기 위해 프롬프트 구조 조정 및 토큰 제한 조절과 불필요 태그 후처리 필터링 등 정제 작업 진행  - 성원선  : 대량 데이터 셋을 전처리(텍스트 길이 조정, 반복 문자 정리, 결측치 제거 등)  한 후 RoBERT를 이용하여 분석을 하였으나 중립 의견이 분석이 제대로 되지않아 3-class 특화 감정분석 모델 & DistilBERT를 이용하여 재시도  - 홍석우  : kaggle 데이터 instacart market basket analysis 를 이용해서 다양한 시각으로 EDA과정을 진행함 외부 데이터 감정,날씨를 같이 이용 할 수 있는 방향을 모색함(kaggle 데이터 이용 사유는 고객 데이터가 포함되어 있어 EDA 과정을 다채롭게 진행 하기 위함)  - 황진웅  : 데이터랩과 마켓컬리의 상품정보 그리고 기후 데이터로 학습 데이터와 타겟값 설정하였으나, 데이터랩이 상대적인 값이라서 해당 시스템 구현은 중지하였음  kaggle 데이터 [Food.com](http://food.com) Recipes and Interactions을 EDA 작업하고 있음 | **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 표현보다 긍정, 부정 데이터가 많이 있음을 확인  - 해시태그 자동 생성 EDA 및 초기구축  **1.1) 자동 평점 시스템**  [전처리작업]  **-** 긍정 부정 감정을 판단내리지못하는 중립적인 표현들 전처리.  전체 평점에 공통적으로 들어가는 단어들 파악하여 전처리.  [EDA작업을 더 정교하게 진행]  -> 실제로 중립표현의 키워드의 종류를 파악  조금더 심층분석하여 중립적인 표현을 아예 빼버리고 이진분류로 긍부정 나누어 학습시킬지 아니면 중립적인 표현을 살릴지 심층 분석하기.  결과가 나오면 모델 학습까지 진행 완료.  회귀로 자동평점 예측시스템 완성.  **1.2) 키워드 자동생성 시스템**  [키워드 분류모델 생성]  - 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**  - 데이터 수집(kaggle 데이터셋, 카카오맵)  - 목표 설정(사용자 취향 및 상태와 테마 기반 음식 추천)  - 대안모색(데이터셋을 구할 수 없는 기능일 시)  **2.1) EDA**  - 수집한 데이터 확인  -> 데이터들의 연관성 분석하기  진행하려는 기능에 필요한 데이터  의미있는 데이터 사용하기,  의미없는 데이터는 버리기.  - 적용할 수 있는 기능 확인  - 추천 시스템에 사용할 수 있는 피쳐들의 상관관계와 외부 데이터를 결합해보기  **2.2) 데이터 수집**  - 지역별 음식점 데이터  - 음식 선호도 관련 데이터 |

**2. 강사님 피드백**

| 중립이란 감정 특성상, 긍/부정 모델과는 전혀 다른 방향에서 다뤄져야 합니다. 현재 Word2Vec로 중립을 파악하려는 접근은 기준이 명확하지 않습니다.  이러한 문제들은 설명이 더 필요하면 말씀해주시고, 데이터 라벨링 기준부터 재정립할 필요가 있습니다.  키워드 기반 작업 이외도 감정 기반 툴로 감정 점수를 수치화하는 방법도 있습니다.  회귀 모델과 분류 모델의 성능 비교가 명확하게 이루어졌으면 좋을 것 같고, 회귀에서는 분류 이후에 정확도를 보는 것이 더 직관적일 수 있습니다.  키워드 자동 생성 시스템의 경우 처리 결과가 궁금하네요. 또한 평가 지표나 품질 향상 방향성이 마련이 되어있는지 확인이 필요합니다.  맞춤형 추천 시스템은 실제 결과에서 사용자에게 의미있는 추천이 될 수 있을지 고민하면서 작업해주시면 좋을 것 같습니다. (현재 작업하는 데이터가 니즈에 맞을지, 모델의 출력 결과가 의미가 있을지) |
| --- |