

SK네트웍스 Family AI 과정 14기  
 데이터 전처리 인공지능 학습 결과서



| 산출물 단계 | 데이터 전처리 |
| --- | --- |
| 평가 산출물 | 인공지능 학습 결과서 |
| 제출 일자 | 2025-09-26 |
| 깃허브 경로 | https://github.com/skn-ai14-250409/SKN14-Final-2Team |
| 작성 팀원 | 한성규 |

**1. 모델 비교 및 선정 이유**

**1.1 실험 개요**

* 문제: 영어 description으로부터 다중 라벨(main\_accord) 예측
* 모델 계열: 다국어 임베딩 모델 + 머신러닝 앙상블 분류기
* 최종 채택 모델: **SentenceTransformer(MiniLM) + VotingClassifier(Soft Voting)**
  + 임베딩: paraphrase-multilingual-MiniLM-L12-v2
  + 분류기: VotingClassifier (앙상블)
    - 기반 모델 1: LogisticRegression
    - 기반 모델 2: XGBClassifier

**1.2 내부 비교**

| 실험 항목 | 설정 | 결과 요약 |
| --- | --- | --- |
| 손실함수 | TF-IDF + LogisticRegression,  **Embedding + Classifier** | 구현이 간단/속도가 빠르나, 문맥적 의미 파악 한계.  **사전 학습된 다국어 임베딩 모델을 통해 향후 다른 언어 입력에 대한 확장성을 확보할 수 있어 채택.** |
|
| 임베딩 모델 | LaBSE, xlm-roberta-base,  MiniLM 등 | MiniLM  (paraphrase-multilingual-MiniLM-L12-v2)이 문장 유사도 기반 향 분류에 가장 적합한 성능을 보임 |
| 분류기 | Logistic Regression, XGBoost, Voting/Stacking | 단일 모델보다 두 모델의 예측 확률을 평균내는  **Soft Voting 앙상블 방식**이 더 안정적이고  성능이 우수함. |
| 데이터 정제 | 희소 라벨  (등장 빈도 10 이하) 및  노이즈 라벨($$$, of 등) 제거 | 모델 학습에 방해가 되는 불필요한 라벨을 제거하여 예측 성능 및 안정성 향상. |
| 전처리 | 브랜드·향수명·  불용어 등 제거 | 문맥 정보 손실로 성능 하락 → **미적용으로 확정** |
| 임계값 튜닝 | 라벨별 최적 임계값 탐색 | 전체 고정 임계값 대신, 각 라벨별로  최적의 예측 임계값을 탐색하여 적용함으로써  **Micro-F1 점수 개선** **(≈0.51).** |

**최종 선정 근거**

* **다국어 처리 확장성**: 영어 데이터로 학습했지만, 다국어 임베딩 모델(MiniLM)을 사용하여 향후 한국어 등 다른 언어의 입력값도 처리할 수 있는 기반을 마련함
* **성능과 효율성**: 파인튜닝 기반의 단일 딥러닝 모델 대비, 사전학습된 경량 임베딩 모델과 머신러닝 분류기의 조합으로 리소스 사용을 최적화.
* **안정성**: VotingClassifier 앙상블을 통해 단일 모델의 편향을 줄이고 더 안정적인 예측 결과를 도출. 내부 실험 결과 Micro-F1 ≈0.51으로 목표 성능 달성.

**2. 모델 구조 및 아키텍처**

* 입력: 영어 description (자연어 텍스트)
* 임베딩: paraphrase-multilingual-MiniLM-L12-v2 모델을 사용하여 텍스트를 고차원 벡터로 변환 (Sentence Embedding).
* 분류: VotingClassifier가 임베딩 벡터를 입력받아 각 라벨에 대한 예측 확률을 출력. 여러 개의 기본 추정기(Estimators)를 사용한 소프트 보팅(Soft Voting) 방식.
* 출력: 예측된 영어 라벨에 대한 17차원 확률 벡터(각 라벨 독립 판정)

**도식 :** Text → SentenceTransformer(MiniLM) → Embedding Vector →

VotingClassifier (LR + XGB) → Probabilities

**3. 학습 설정 및 하이퍼파라미터**

| 항목 | 값 |
| --- | --- |
| 학습/검증 데이터 분할 | 총 26,319 중  빈 라벨 264 제외 후26,055  80% / 20% |
| 희소 라벨 제거 기준 | RARE\_MIN\_COUNT =10  (등장 빈도 10회 이하) |
| 노이즈 라벨 | '$', 'of', 'the' |
| 임베딩 모델 | paraphrase-multilingual-MiniLM-L12-v2 |
| 분류 모델 | OneVsRestClassifier(VotingClassifier) |
| Voting 앙상블 구성 | LogisticRegression(max\_iter=200), XGBClassifier |
| 예측 임계값 (Threshold) | 라벨별 최적화 (0.2 ~ 0.5 범위에서 탐색) |
| Top-K 예측 (후보) | 3 |
| Token max length | 256 |
| 배치 크기 (Batch Size) | 16 |

* 저장 방식: 모델이 딥러닝(임베딩)과 머신러닝(분류기)으로 분리되어 있어, 각기 다른 형식으로 저장.
  + MiniLM 임베딩 모델: minilm\_model.pt (PyTorch)
  + 분류기 및 라벨 정보: label\_info.pkl (Scikit-learn)

**4. 학습 결과 및 성능 평가**

**4.1 주요 지표(검증 셋)**

* 라벨별 임계값 최적화 적용 시 최고점
  + Micro-F1 ≈ 0.5001
  + Macro-F1 ≈ 0.2513
  + Sample-F1 ≈ 0.4985
* Micro-F1: 전체 샘플에 대한 TP, FP, FN을 종합하여 계산. 데이터 내 다수 클래스(Head labels)의 성능을 잘 나타내어 전반적인 모델 성능 지표로 활용.
* Macro-F1 / Sample-F1: 소수 클래스(Tail labels)에 대한 성능 및 사용자 체감 성능을 나타내는 지표로, 라벨 불균형 문제로 인해 상대적으로 낮게 측정되었음.

### 4.2 전처리 영향

* 향을 묘사하는 핵심 단어 외에 문맥을 구성하는 전치사, 접속사 등의 정보가 중요하게 작용.
* 불용어 및 브랜드명 등 텍스트 정규화 과정은 오히려 문맥 정보를 손실시켜 최종 성능을 하락시키는 것으로 확인되어 적용하지 않음.

**5. 과적합/과소적합 대응**

| 기법 | 설명 | 적용 여부 |
| --- | --- | --- |
| 데이터 정제 | 희소 라벨 및 노이즈 라벨 제거 | O |
| 정규화 | 텍스트 정리(브랜드/향수명/숫자 제거) | X |
| 앙상블 | VotingClassifier로 여러 모델의 예측을 종합하여 일반화 성능 향상 | O |
| 임계값 튜닝 | 라벨별 최적 threshold 탐색 및 저장 | O |

**6. 결론 및 향후 계획**

* **결론**: 다국어 처리가 가능한 SentenceTransformer(MiniLM) 임베딩과 VotingClassifier 앙상블을 결합한 모델을 최종 채택하여 Micro-F1 약 0.5001의 성능을 달성함. 이는 경량 모델이면서도 한국어 등 다른 언어에 대한 확장성을 확보했다는 점에서 의미가 있음.
* **향후 계획**
  1. 데이터 보강: 더 많은 향수 리뷰 및 설명 데이터를 수집하여 라벨링 정확도와 데이터 수를 늘려 모델 재학습.
  2. 임계값 최적화: 고정 임계값이 아닌, 각 라벨의 특성을 고려한 개별 최적 임계값을 탐색하여 적용.
  3. 피처 엔지니어링: 텍스트 외 추가적인 메타데이터(브랜드, 성별 등)를 활용하여 성능 개선.