

**SK네트웍스 Family AI 과정 15기  
 데이터 전처리 인공지능 학습 결과서**



| **산출물 단계** | 데이터 전처리 |
| --- | --- |
| **평가 산출물** | 인공지능 학습 결과서 |
| **제출 일자** |  |
| **깃허브 경로** | 깃허브 주소 기재 |
| **작성 팀원** |  |

**1. 모델 비교 및 선정 이유**

* 비교 대상 모델:

| **모델명** | **종류** | **선정 이유** |
| --- | --- | --- |
| BERT | Transformer 기반 사전학습 모델 | 문장 의미 파악에 강함, 사전학습 모델 활용 가능 |
| BiLSTM | RNN 변형 | 시퀀스 특성 반영에 적합, 구조 간단 |

* 실험 모델 수: 총 2종
* 최종 선정 모델: 예) BERT

**2. 모델 구조 및 아키텍처**

2.1 모델 아키텍처 도식 (선택사항: 도식 첨부 또는 말로 설명)

* 입력층 → 임베딩층 → 인코더블록(N개) → 분류기 출력층 (Softmax)

2.2 구성 요소 설명:

| **계층명** | **역할** | **구성 요소** |
| --- | --- | --- |
| Embedding | 입력 문장을 벡터화 | WordPiece tokenizer 사용 |
| Encoder | 의미 표현 학습 | Multi-Head Attention + FFN |
| Classification Head | 태그 예측 | Dense Layer + Softmax |

**3. 학습 설정 및 하이퍼파라미터**

| **항목** | **값** |
| --- | --- |
| 학습 데이터 수 | 10,000건 |
| 검증 데이터 수 | 2,500건 |
| 에폭(Epoch) 수 | 10 |
| 배치 크기 (Batch Size) | 32 |
| 학습률 (Learning Rate) | 2e-5 |
| 옵티마이저 | AdamW |
| 손실 함수 | CrossEntropyLoss |
| 조기 종료 기준 | validation loss 3회 이상 증가 시 종료 |

**4. 학습 결과 및 성능 평가**

4.1. 학습 결과 요약

| **모델** | **Accuracy** | **Precision** | **Recall** | **F1 Score** |
| --- | --- | --- | --- | --- |
| BERT | 0.914 | 0.92 | 0.91 | 0.915 |
| BiLSTM | 0.872 | 0.87 | 0.86 | 0.865 |

4.2. 그래프 (선택): 학습/검증 loss 변화, accuracy 변화 등

(예: matplotlib 또는 캡처 첨부)

4.3 해석 및 분석

* F1 Score가 가장 높은 모델: BERT
* 특정 클래스 오분류 경향: “날씨” ↔ “일상정보” 유사 문장 혼동

**5. 과적합/과소적합 대응**

* 적용 기법:

| **기법** | **설명** | **적용 여부** |
| --- | --- | --- |
| Dropout | 과적합 방지 | O (0.1 사용) |
| 조기 종료 (Early Stopping) | 성능 저하 시 종료 | O |
| 학습률 감소 | Plateau 시 자동 감소 | O (ReduceLROnPlateau 사용) |
| 교차 검증 | 데이터 분산 고려 | X (단일 validation set 사용) |

* 결과: 학습/검증 간 loss 차이가 적고 성능 안정

6. 결론 및 향후 계획

* 최종 선정 모델: BERT
* 활용 방안: 사용자 챗봇 응답 분류기, 유사 질문 추천
* 향후 계획:  
  + 테스트 데이터로 실사용 환경 평가 예정
  + 모델 경량화(TinyBERT 또는 distillation) 고려
  + 웹 연동 및 REST API 서비스화

7. 부록

* 전체 학습 로그 캡처 또는 파일 (예: TensorBoard, wandb 스크린샷)
* 학습 코드 경로 또는 GitHub 링크
* 주요 파라미터 설정 config.json 또는 코드 스니펫