**NLP 주요 평가 지표 설명**

**1. 개요**

모델 성능 평가에 사용되는 주요 지표인 Precision, Recall, F1 Score, Accuracy의 수식, 의미, 간단한 예시를 정리함.

**2. 정확도 지표 수식 및 설명**

|  |  |  |
| --- | --- | --- |
| 지표 | 수식 | 의미 |
| Precision (정밀도) | TP / (TP + FP) | 모델이 "정답"이라고 예측한 것 중 실제 정답의 비율임 |
| Recall (재현율) | TP / (TP + FN) | 실제 정답 중 모델이 찾아낸 비율임 |
| F1 Score | 2 × (Precision × Recall) / (Precision + Recall) | Precision과 Recall의 조화 평균임 |
| Accuracy (정확도) | (TP + TN) / (TP + FP + TN + FN) | 전체 데이터 중 맞춘 정답 비율임 |

※ TP: True Positive, FP: False Positive, TN: True Negative, FN: False Negative를 의미함.

**3. 간단한 수치 예시**

**가정**

* 총 데이터 수: 100개
* 실제 긍정 데이터 수: 20개
* 모델 예측 결과:
  + True Positive(TP): 10개
  + False Positive(FP): 5개
  + True Negative(TN): 70개
  + False Negative(FN): 10개

**계산 결과**

|  |  |  |
| --- | --- | --- |
| 지표 | 계산 과정 | 결과 |
| Precision | 10 / (10 + 5) = 10/15 | 0.6667 (66.7%) |
| Recall | 10 / (10 + 10) = 10/20 | 0.5 (50.0%) |
| F1 Score | 2 × (0.6667 × 0.5) / (0.6667 + 0.5) ≈ 0.5714 | 57.1% |
| Accuracy | (10 + 70) / (10 + 5 + 70 + 10) = 80/95 | 0.8421 (84.2%) |

* Precision이 높으면 모델이 예측한 정답 중 실제 맞춘 비율이 높음.
* Recall이 높으면 실제 정답 중 놓치지 않고 잘 찾아냄.
* F1 Score는 Precision과 Recall을 균형 있게 종합한 지표임.
* Accuracy는 전체 중 정답률을 평가하지만, 데이터 불균형 상황에서는 F1 Score가 더 신뢰할 수 있는 평가 지표임.

**4. NLP 지표 수식 및 설명**

|  |  |  |
| --- | --- | --- |
| 지표 | 설명 | 예시 |
| MNLI Multi-Genre Natural Language Inference | 문장(전제)과 문장(가설) 사이 관계를 분류 (Entailment, Contradiction, Neutral) | **전제**: "The dog is running in the park." **가설**: "An animal is outside." → **Entailment (함의)** |
| QQP Quora Question Pairs | 두 질문이 같은 의미인지 판별 | "How do I bake a cake?" "What are the steps to make a cake?" → **Paraphrase (의미 같음)** |
| QNLI Question Natural Language Inference | 문장이 질문에 대한 답을 포함하는지 판별 | **질문**: "Where is the Eiffel Tower located?" **문장**: "The Eiffel Tower is in Paris." → **Yes (포함함)** |
| SST-2 Stanford Sentiment Treebank | 문장의 감정(긍정/부정) 분류 | "The movie was absolutely wonderful." → **Positive (긍정)** |
| CoLA Corpus of Linguistic Acceptability | 문장이 문법적으로 맞는지 판별 | "She has visited Paris." → **Grammatically acceptable (맞음)** |
| STS-B Semantic Textual Similarity Benchmark | 두 문장이 얼마나 비슷한지 점수 매김 (0~5) | "A man is playing a guitar." "Someone is strumming a musical instrument." → **4.5 (매우 비슷함)** |
| MRPC Microsoft Research Paraphrase Corpus | 두 문장이 의미적으로 같은지 판단 | "He canceled the meeting." "The meeting was called off by him." → **Paraphrase (의미 같음)** |
| RTE Recognizing Textual Entailment | 두 문장 사이에 함의 관계가 있는지 판별 | **문장1**: "A man is riding a horse." **문장2**: "Someone is on an animal." → **Entailment (함의)** |