

De-identification Techniques

Safe Harbor Method

18가지 식별자를 제거하는 HIPAA 표준 방법

- 이름, 주소, 날짜 등 명시된 항목 제거
- 구현이 상대적으로 간단
- 규정 준수 용이

Expert Determination

전문가가 재식별 위험을 매우 낮은 수준으로 판단

- 통계적 방법 활용
- 더 많은 데이터 활용 가능
- 전문가 검증 필요

Rule-based Pattern Matching

정규표현식을 사용한 자동 탐지

- 날짜 패턴: $\backslash d\{2\} \backslash d\{2\} \backslash d\{4\}$
- 전화번호: $\backslash d\{3\} - \backslash d\{3\} - \backslash d\{4\}$
- 빠른 처리 속도

ML-based Detection

머신러닝 기반 NER 모델 활용

- BiLSTM-CRF, BERT 모델
- 문맥 기반 탐지 가능
- F1 score 95%+ 달성

정확도 메트릭 비교

Precision

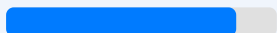
정밀도



탐지된 항목 중 실제 PHI 비율

Recall

재현율



실제 PHI 중 탐지된 비율

F1 Score

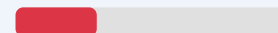
조화 평균



Precision과 Recall의 조화 평균

FP Rate

오탐률



비PHI가 PHI로 잘못 탐지