De-identification Techniques

Safe Harbor Method

HIPAA standard method that removes 18 identifiers

- Removes specified items such as names, addresses, dates
- Relatively simple implementation
- Easy regulatory compliance

Expert Determination

Expert judges re-identification risk to be very low

- Utilizes statistical methods
- Allows use of more data
- Requires expert verification

Rule-based Pattern Matching

Automatic detection using regular expressions

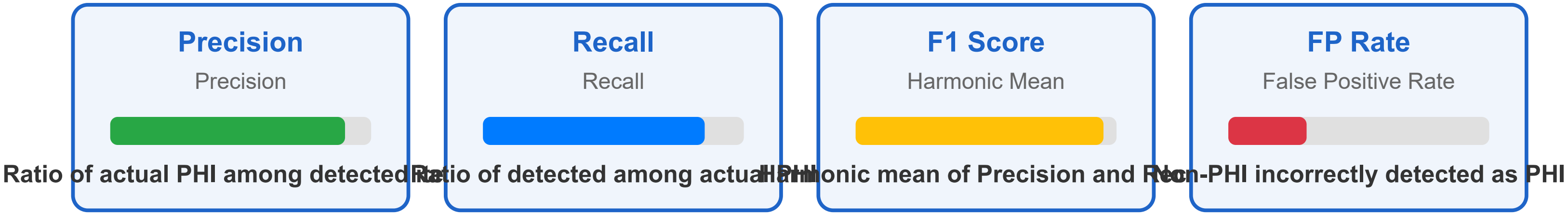
- Date pattern: \d{2}/\d{2}/\d{4}
- Phone number: \d{3}-\d{3}-\d{4}
- Fast processing speed

ML-based Detection

Utilizing machine learning-based NER models

- BiLSTM-CRF, BERT models
- Context-based detection possible
- Achieves F1 score of 95%+

Accuracy Metrics Comparison



Precision

Precision

Ratio of actual PHI among detected

Recall

Recall

Ratio of detected among actual PHI

F1 Score

Harmonic Mean

Harmonic mean of Precision and Recall

FP Rate

False Positive Rate

Non-PHI incorrectly detected as PHI