USC University of Southern California

USC Viterbi School of Engineering

Reading Report 2

Deep learning-based NLP data pipeline for EHR-scanned document information extraction

Introduction

- Challenge:

  Scanned EHR documents pose a significant challenge due to their image format, which complicates the extraction of vital patient information.

- Opportunity:

  By utilizing deep learning and NLP techniques, healthcare data management and decision-making can be significantly enhanced, providing valuable insights from these documents.

- Objective:

  The goal is to develop a robust data pipeline that leverages image preprocessing, Optical Character Recognition (OCR), and advanced NLP models to extract critical medical details from scanned sleep study reports.

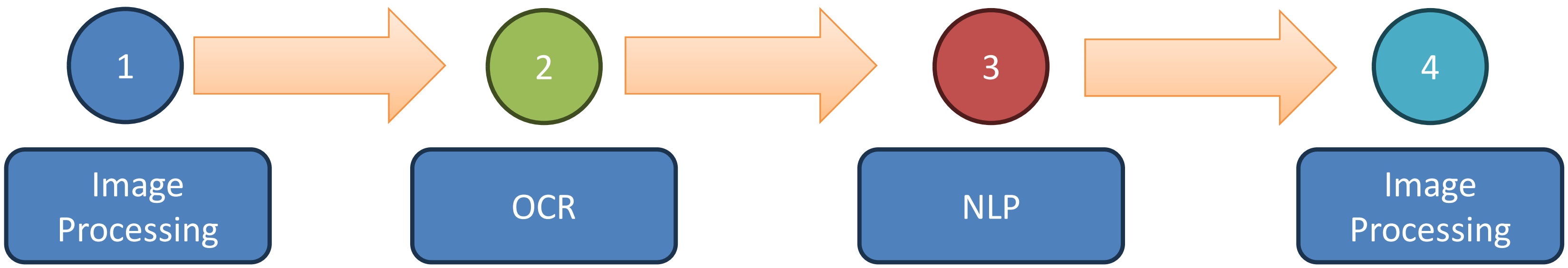
- Data Focus:

  The primary focus is on crucial sleep apnea indicators, including:

  AHI (Apnea-Hypopnea Index): A measure of sleep apnea severity.

  SaO2 (Oxygen Saturation): Provides essential clinical insights.

Information Extraction Pipeline



- Transforms scanned document images into formats optimized for OCR
- Techniques: Grayscaling, dilating, eroding, adjusting contrast

- Transforms scanned documents into a readable format for further analysis
- Utilizes Tesseract to convert images into text

- Models: Traditional machine learning and deep learning-based NLP models (e.g., bag-of-words)
- Transformer-based deep learning models (e.g., ClinicalBERT)

- Tested: Structured and unstructured input formats
- Analyzed: Context and accuracy

Data and Image Processing

- Data Source:
  - Origin: UTMB EHR reports
  - Dataset: 2988 scanned PDFs from 955 unique reports
- Image Processing:
  - Techniques: Convert to grayscale, apply dilation and erosion, increase contrast using OpenCV
- OCR:
  - Tool: Tesseract OCR via pytesseract for text extraction
  - Validation: Visual inspection of extracted text
- Deidentification:
  - Process: Masked patient names, medical record numbers, and dates
- Text Segmentation & Classification:
  - Segmentation: Identify AHI and SaO2 values with context
  - Classification: Bag-of-words, BiLSTM, BERT, ClinicalBERT
- Model Training and Evaluation
  - Training: 70:30 sets; cross-validation, checkpoint-based evaluation
  - Metrics: Recall, precision, AUROC, document accuracy
- Additional Analysis
  - Training Set Size: Effect on model performance
  - Validation: Impact of preprocessing methods and feature contribution

USC Viterbi School of Engineering

Results

- Report Formats: Varied with text, images, and handwriting
- OCR Issues: Text extraction mostly successful; challenges with images and handwriting
- Data: Median of 2 pages, 44 numeric values per page. AHI avg. 34.9, SaO2 avg. 76.5
- Model Performance: ClinicalBERT outperformed others in AHI and SaO2 extraction
- Training Size: ClinicalBERT excelled with fewer reports; all models performed similarly with 50 reports
- Preprocessing: Best results with contrast increase and structured input

Quiz Questions

- **Question:** Why is ClinicalBERT advantageous over traditional machine learning models?
- **Answer:**
  - Due to its bidirectional transformer architecture, it was pre-trained on 2 million clinical notes. This helps in enabling better comprehension of context