Research Document: EHR Data Cleansing with NLP and Machine Learning

Abstract

This project demonstrates an automated approach to cleansing and standardizing Electronic Health Records (EHRs) using Natural Language Processing (NLP) and machine learning. The pipeline addresses common data quality issues in healthcare datasets, such as inconsistent terminology, typographical errors, and non-standard date formats, to enable more reliable downstream analytics.

Introduction

Electronic Health Records are a rich source of patient information but are often plagued by inconsistencies and errors. This project aims to automate the cleansing of EHR data using a combination of NLP, fuzzy matching, and machine learning techniques, improving data quality for research and clinical decision-making.

Data Description

The dataset consists of 55,500 records with 15 fields, including:

- Name, Age, Gender, Blood Type
- Medical Condition, Date of Admission, Discharge Date
- Doctor, Hospital, Insurance Provider
- Billing Amount, Room Number, Admission Type
- Medication, Test Results

Sample record:

```
Brooke Brady,44,Female,AB+,Cancer,2021-10-08,Roberta Stewart,Morris-Arellano,UnitedHealthcare,40701.60,182,Urgent,2021-10-13,Paraceta
```

Methodology

The data cleansing pipeline consists of the following steps:

1. **Name Standardization**: Proper capitalization, removal of extra spaces, and handling of hyphenated names.
2. **Medical Term Standardization**: Expansion of common medical abbreviations (e.g., HTN → Hypertension) and title-casing using spaCy and regex.
3. **Misspelling Correction**: Fuzzy matching (RapidFuzz) to correct typographical errors in medical conditions and medication names.
4. **Date Format Standardization**: Parsing and formatting all date fields to ISO format (YYYY-MM-DD) using `dateutil`.
5. **Data Quality Analysis**: Quantitative and visual analysis of the impact of standardization on unique values and distributions.

Tools and Libraries

- pandas, numpy: Data manipulation
- spaCy: NLP and entity recognition
- RapidFuzz, FuzzyWuzzy: Fuzzy string matching
- matplotlib, seaborn: Visualization
- dateutil: Date parsing
- icd10-cm: ICD-10 code standardization

Results

- **Names standardized**: 55,467 out of 55,500 records
- **Medical Conditions (unique)**: Reduced to 6 standardized terms
- **Medications (unique)**: Reduced to 5 standardized terms
- **Date formats**: All admission and discharge dates standardized to ISO format

Example Changes

Original Name	Standardized Name	Original Condition	Standardized Condition
Bobby JacksOn	Bobby Jackson	Cancer	Cancer
LesLie TErRy	Leslie Terry	Obesity	Obesity
DaNnY sMitH	Danny Smith	Obesity	Obesity

Original Name	Standardized Name	Original Condition	Standardized Condition
andrEw waTtS	Andrew Watts	Diabetes	Diabetes
adrIENNE bEll	Adrienne Bell	Cancer	Cancer

Top Standardized Medical Conditions

Arthritis:	9,308
Diabetes:	9,304
Hypertension:	9,245
Obesity:	9,231
Cancer:	9,227
Asthma:	9,185

Top Standardized Medications

Lipitor:	11,140
Ibuprofen:	11,127
Aspirin:	11,094
Paracetamol:	11,071
Penicillin:	11,068

Discussion

The automated pipeline significantly improved data consistency, reducing the number of unique terms and correcting thousands of typographical errors. Standardizing terminology and dates enables more accurate analytics and research. The approach is extensible to other healthcare datasets and can be customized with domain-specific dictionaries.

Conclusion

Automated EHR data cleansing using NLP and machine learning is effective for improving data quality. This project provides a reproducible pipeline that can be adapted for various healthcare data sources, facilitating better research and clinical outcomes.

References

- spaCy: https://spacy.io/ (https://spacy.io/)
- RapidFuzz: https://maxbachmann.github.io/RapidFuzz/ (https://maxbachmann.github.io/RapidFuzz/)
- FuzzyWuzzy: https://github.com/seatgeek/fuzzywuzzy (https://github.com/seatgeek/fuzzywuzzy)
- ICD-10-CM: https://www.cdc.gov/nchs/icd/icd10cm.htm (https://www.cdc.gov/nchs/icd/icd10cm.htm)
- pandas: https://pandas.pydata.org/ (https://pandas.pydata.org/)
- scispaCy: https://allenai.github.io/scispacy/ (https://allenai.github.io/scispacy/)