Reproducibility Report: Bi-Dimensional Representation of Patients for Diagnosis Prediction

Anonymous submission

Abstract

This report presents a reproducibility study on the paper "Bi-Dimensional Representation of Patients for Diagnosis Prediction." We implement multiple BiLSTM-based models with different symptom vectorization techniques (TF-IDF, Word2Vec, and Hybrid) to classify ICD-9 codes from clinical discharge summaries. We evaluate the reproducibility of the paper's methodology, highlight encountered challenges, and summarize experimental results.

Project Summary

Objective: Reproduce and evaluate models for ICD-9 code classification from discharge summaries, exploring different feature extraction techniques.

Approach:

- **Data Preprocessing**: Extracted and cleaned discharge summaries, focusing on relevant clinical sections.
- **Symptom Extraction**: Utilized SciSpaCy with UMLS linker to extract symptoms, targeting semantic type "T184."
- **Modeling**: Implemented BiLSTM models with three vectorization strategies: TF-IDF, Word2Vec, and a hybrid of both.

Experiments Conducted

Model Variants:

- TF-IDF + BiLSTM: Converted symptom lists into TF-IDF vectors and trained a BiLSTM model.
- Word2Vec + BiLSTM: Trained Word2Vec embeddings on symptom tokens and used them as input to the BiLSTM.
- Hybrid (TF-IDF + Word2Vec) + BiLSTM: Combined TF-IDF and Word2Vec features for each symptom token and trained the BiLSTM.

Evaluation Metrics:

- Accuracy
- Macro and Micro F1-scores
- Confusion Matrix
- ROC AUC Score

Reproducibility Assessment

Challenges:

- **Absence of MetaMap**: Limited the ability to extract comprehensive semantic information and symptoms.
- **Class Imbalance**: The dataset exhibited significant class imbalance, affecting model performance.

Mitigations:

- Focused on the top 50 most frequent ICD-9 codes to reduce complexity.
- Applied class weighting in the loss function to address imbalance.

Outcome: Despite the challenges, the models achieved reasonable performance, indicating partial reproducibility of the original study.

Results Overview

Links to Artifacts

- **Codebase**: `https://github.com/yourusername/yourprojectrepo`
- **Presentation Video**: `https://drive.google.com/yourpresentationlink`

Appendix: Blue Question Responses