CS598 Final Project

Binbin Weng

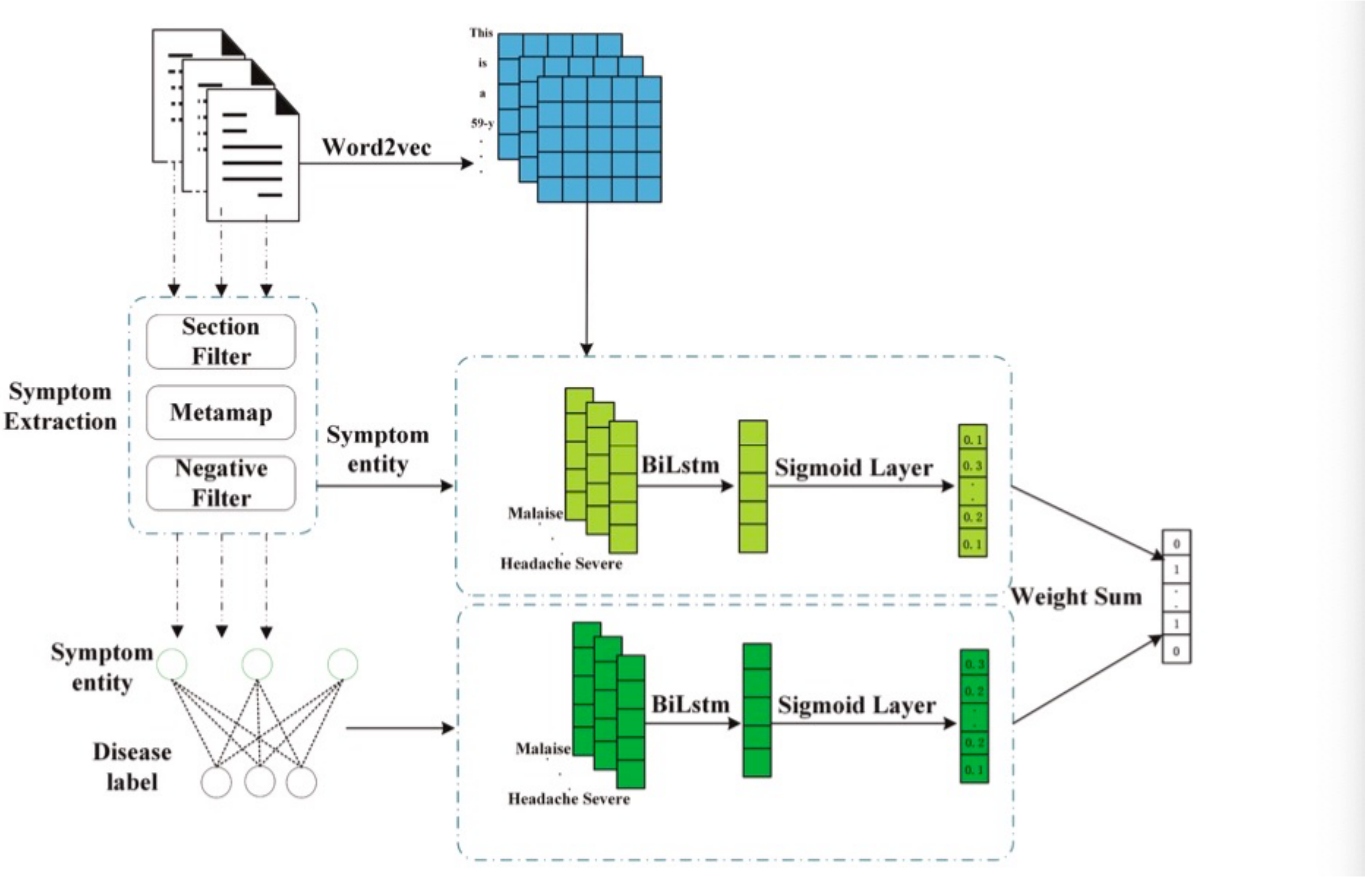
(binbinw2@Illinois.edu)

A disease inference method based on symptom extraction and bidirectional Long Short Term Memory networks

Donglin Guo, Guihua Duan, Ying Yu, Yaohang Li, Fang-Xiang Wu, Min Li

Purpose:

- Data used is clinical text data. Some techniques will be used on the clinical text data to extract symptoms from the text.
- Represent the extracted symptoms in two ways
  - Word2Vec (an Embedding method)
  - TF-IDF (Term Frequency – Inverse Document Frequency)
- Develop a multi-label classifier for disease inference by building two bidirectional Long Short Term networks (BiLSTM) with these two representations of the extracted symptoms to improve the performance of the classifier.



Datasets

- NOTEEVENTS
- DIAGNOSES-ICD

Data Processing

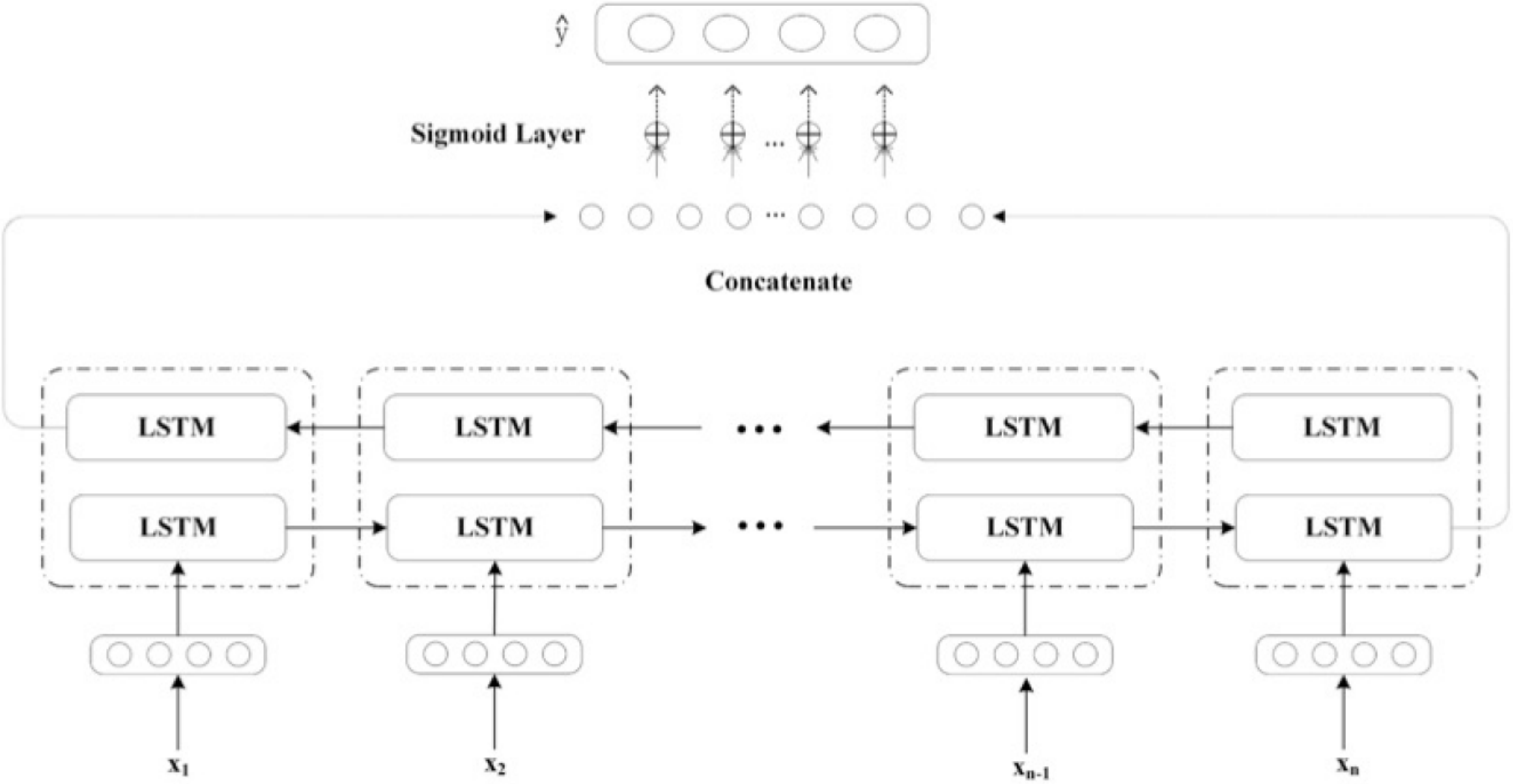
- Merge the two tables with Visit ID
- Prepare clinical text data for symptom extraction
- Use Batch MetaMap service provided by National Library of Medicine to extract symptoms
- Calculate TF-IDF scores for each symptom with each disease
- Build Word2Vec model for symptoms
- Prepare the forward input and backward input for the models

Data Description

- Over 46,000 observations with over 18,000 unique symptoms and 50 unique diseases

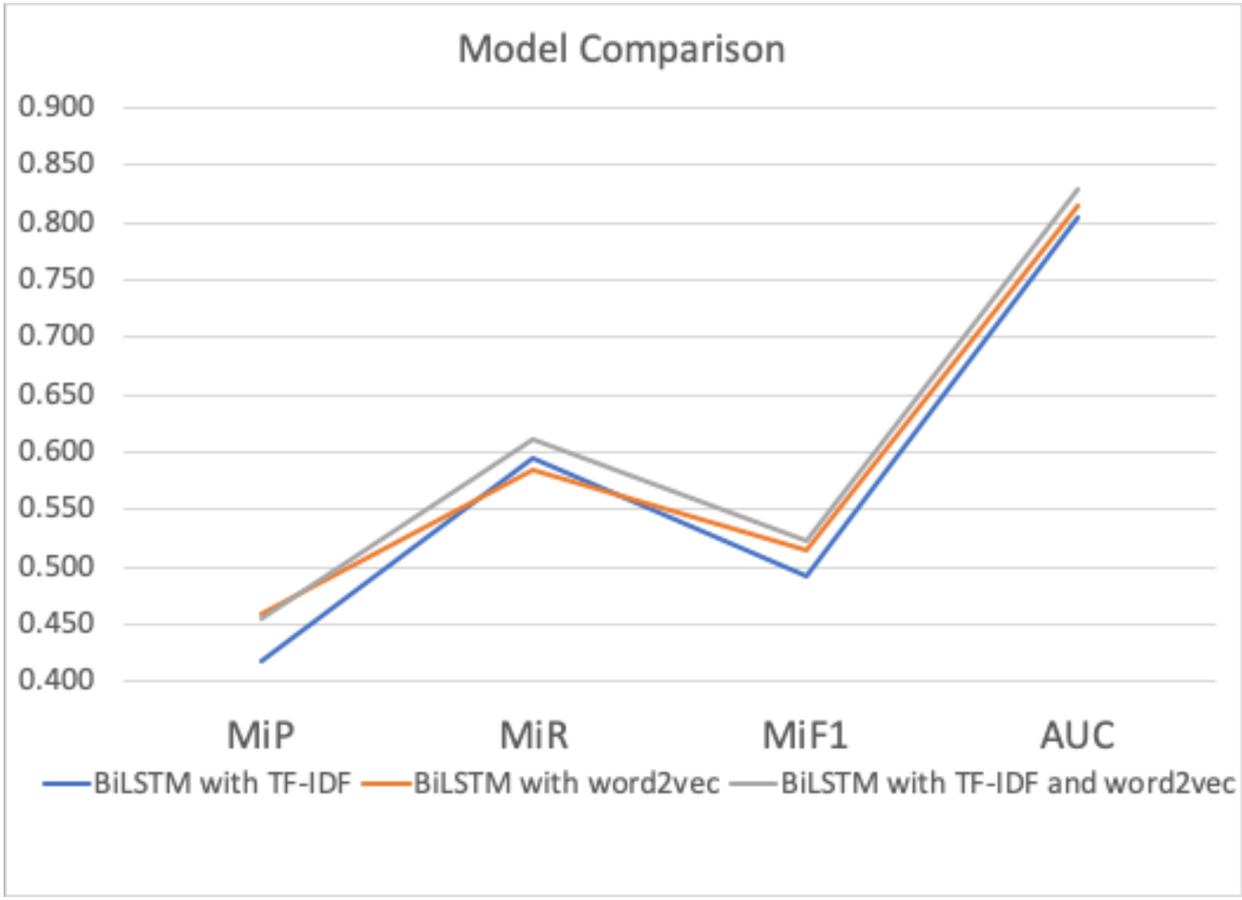
- https://physionet.org/content/mimiciii/1.4/
- https://www.nlm.nih.gov/

Modeling:


$\hat{y}$
Sigmoid Layer
Concatenate
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
x_1
x_2
x_{n-1}
x_n

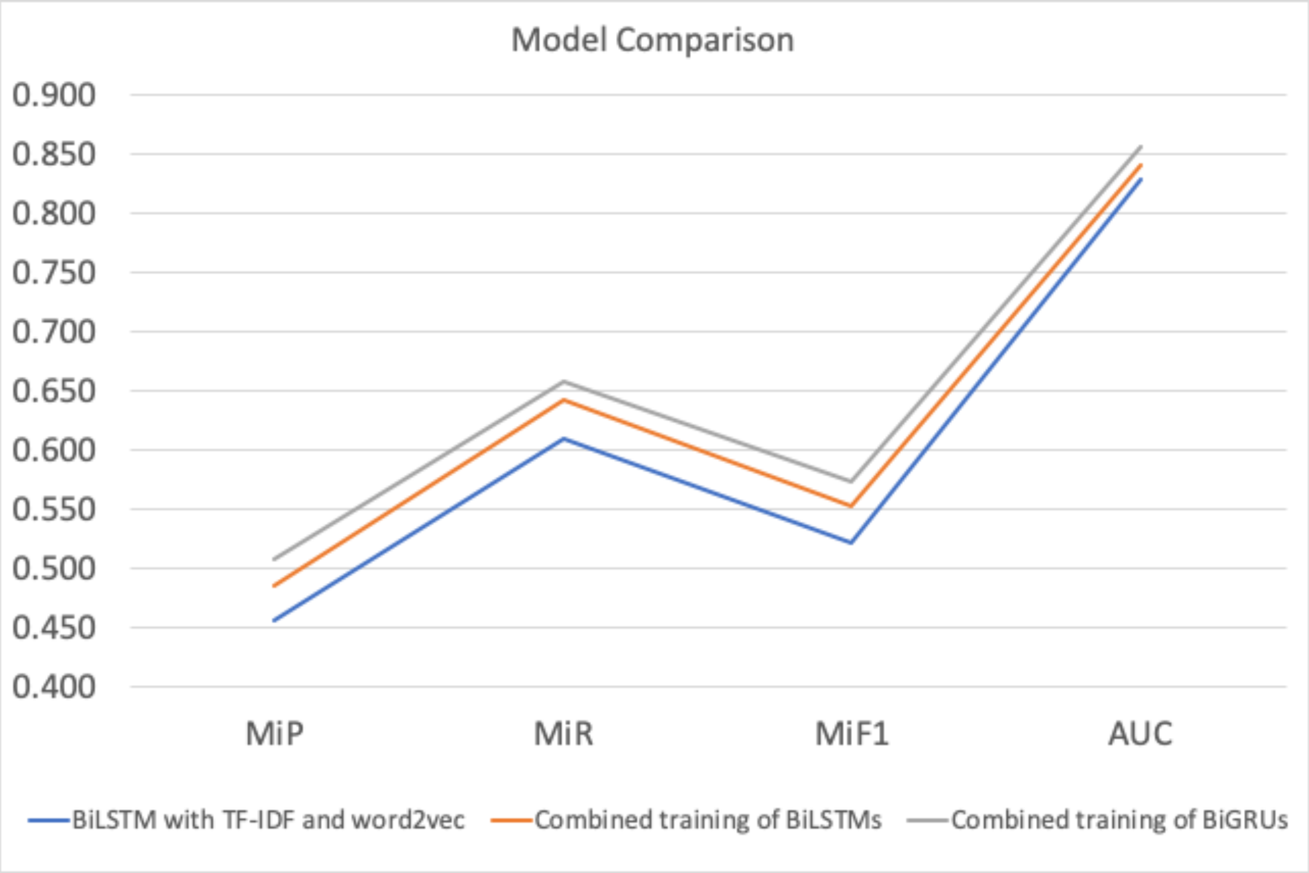
Results

Model	MiP	MiR	MiF1	AUC	Runtim for Trainning
BiLSTM with TF-IDF	0.419	0.596	0.492	0.804	0.681
BiLSTM with word2vec	0.460	0.584	0.515	0.814	0.883
BiLSTM with TF-IDF and word2vec	0.456	0.611	0.522	0.829	1.564



Additional Results

Model	MiP	MiR	MiF1	AUC	Runtim for Trainning
BiLSTM with TF-IDF and word2vec	0.456	0.611	0.522	0.829	1.564
Combined training of BiLSTMs	0.486	0.643	0.554	0.842	1.613
Combined training of BiGRUs	0.508	0.659	0.573	0.856	1.507



Thanks!