ScoreIt-RoBERTa: Scoring clinical patient notes

Ayush Tripathi[1] , Kishan N. Murthy[1] , Paritosh Singh[1] , Prakash Parajuli[1] and Siddharth G. Byale[1]

[1]University of Southern California, Los Angeles, CA, United States of America

Email: ayusht@usc.edu, knarasim@usc.edu, paritosh@usc.edu, parajuli@usc.edu, sbyale@usc.edu

USC Viterbi School of Engineering *Department of Computer Science*

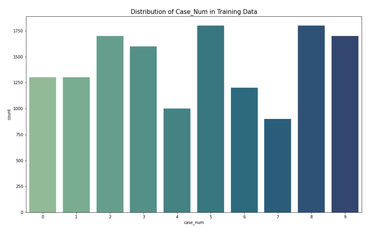
Motivation

Physicians have a long history of writing patient notes that document history, physical exam findings, possible diagnoses, and follow-up care. Extracting meaningful information from these clinical notes can aid in the accurate diagnosis of diseases.

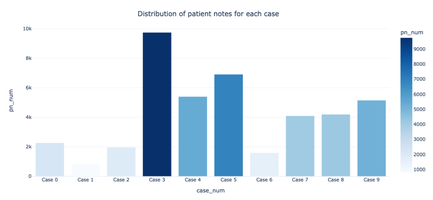
Named Entity Recognition (NER) is a part of information retrieval and natural language processing that tries to recognize and categorize named entities referenced in unstructured text. The objective of our project is to train a NER model capable of identifying and annotating specific clinical concepts in English patient notes. The developed model can also benefit exams like the United States Medical Licensing Examination (USMLE), where extracted features from the clinical notes can be evaluated against the rubric.

In Medical NER, recognition models based on English clinical patient notes are lacking. We propose designing an English NER model that identifies and annotates the text to find specific clinical concepts based on clinical patient notes.

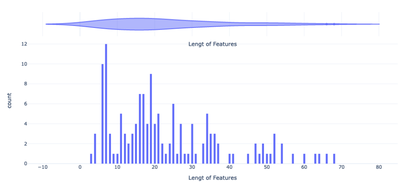
Exploratory Data Analysis



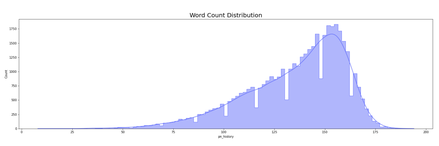
Distribution of Cases in Train Dataset



Distribution of Patient Notes in each case



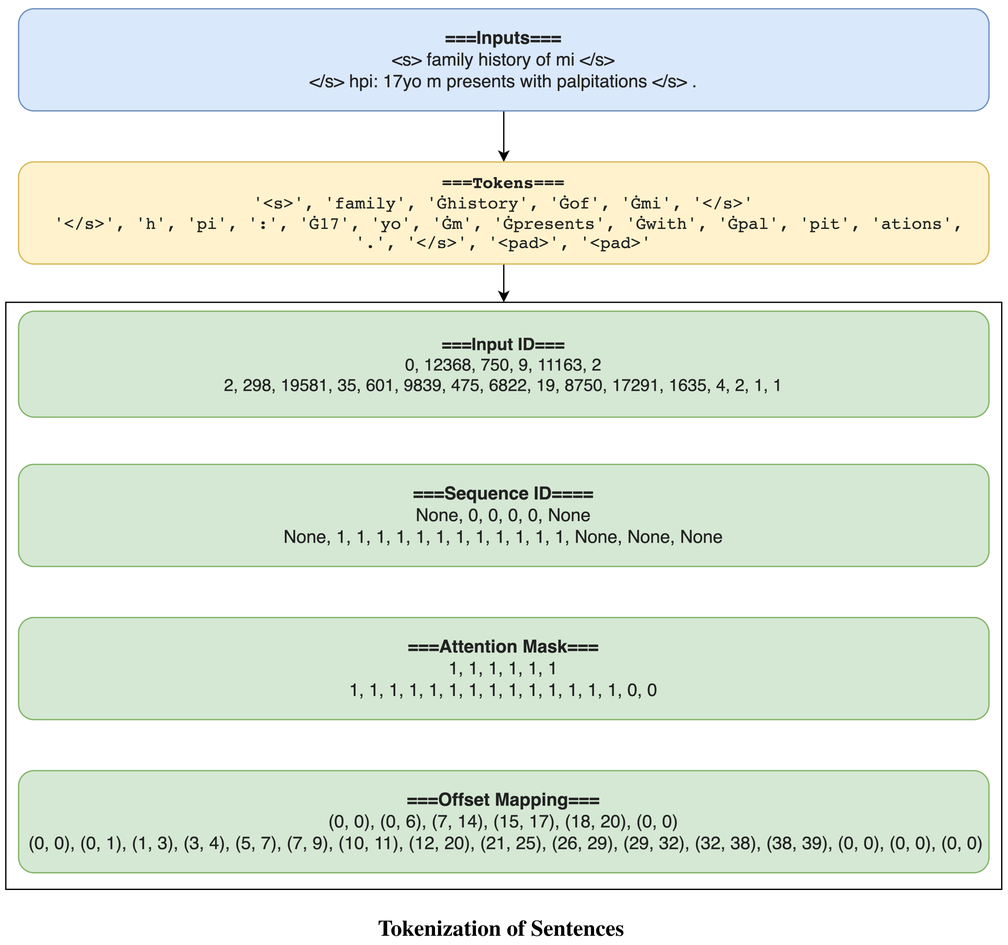
Distribution of Length of Features



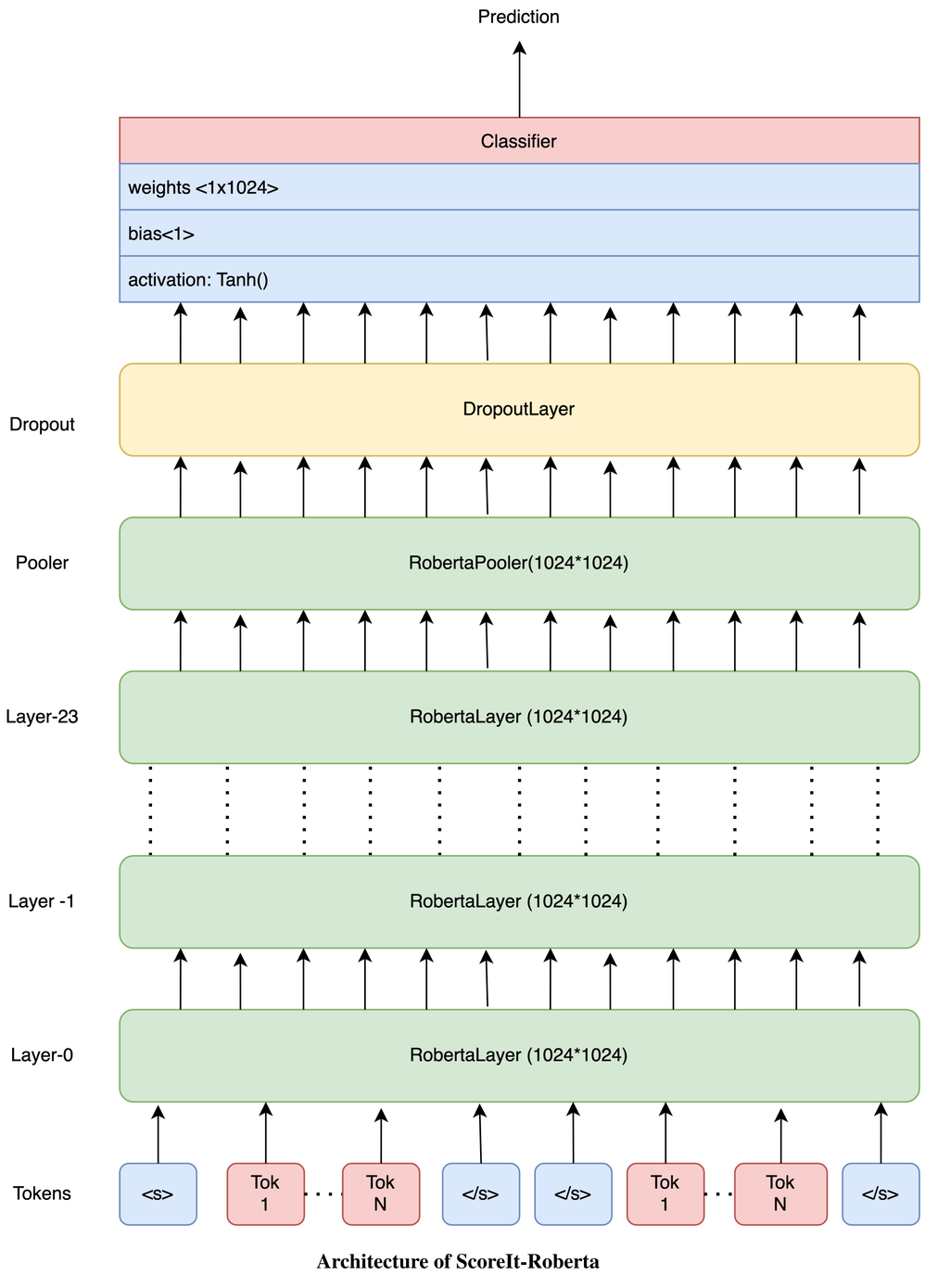
Distribution of Word Count

Architecture

- "RoBERTa-large" model was used for extracting the patient's feature information from the patient history.
- The feature text and the patient history were passed to the tokenizer to get the sequence_ids, input_ids and attention mask. These fields are used as an input to our model.
- The RoBERTa Model consists of 23 layers of RoBERTa, followed by a pooled layer, which has the activation function as Tanh.
- A single layer of RoBERTa is divided into sub-layers, which are the attention layer, output layer, intermediate layer and the output layer.
- At the end of the roberta-model, a dropout layer was added. Finally, a linear classifier was used for the prediction.
- The model was trained on over 355 million parameters.



Tokenization of Sentences



Architecture of ScoreIt-Roberta

Results and Insights

We selected three important indicators: precision, recall and F1 score, to evaluate the performance of the named entity recognition task on the dataset presented from the USMLE Step 2 Clinical Skills examination.

RoBERTa was trained for 5 epochs with a stratified k fold split (k=5), yielding the following results.

Metric	Baseline	BERT	RoBERTa
Precision	96.64	75.27	79.89
Recall	12.99	72.34	84.98
F1-Score	22.91	71.27	82.35

Metrics for Baseline, BERT, RoBERTa

35 yo female Annotation patient presenting for problems with her period for the past 6 months. The patient reports heavy flow and that she needs to change her tampon every 2 hours. Last menstrual period was 2 mo ago. She didn't have similar problems before. Her menstrual cycle is regular. Her first menses occured at the age of 12. No vaginal discharge or blood clots reported. Multiple failed attempts of getting pregnant reported. Her last pap smear was normal 6 months ago. She reports weight gain Annotation gain of 10 pounds Annotation problems with her period Annotation for the past 6 months. The patient reports heavy flow Annotation and that she needs to change her tampon every 2 hours Annotation . Last menstrual period was 2 mo ago Annotation . She didn't have similar problems before. Her menstrual cycle is regular. Her first menses occured at the age of 12. No vaginal discharge or blood clots reported. Multiple failed attempts of getting pregnant reported. Her last pap smear was normal 6 months ago. She reports weight gain of 10 pounds recently with an increased appetite. No cold intolerance, tremors, low mood, constipation, or bloating reported. She reports being anxious at work.

ROS: as listed above

PMH/PSH: none

Medication and allergies: none

FH: negative for spontaneous abortions or heavy bleeds

SH: non smoker, occasional drinker, no illicit drug use, she lives with her 2 adopted daughters, divorced, sexually active with her boyfriend with no contraceptive use Annotation Multiple failed attempts of getting pregnant Annotation 35 yo Annotation female patient presenting for problems with her period for the past 6 months Annotation . The patient reports heavy flow and that she needs to change her tampon every 2 hours. Last menstrual period was 2 mo ago. She didn't have similar problems before. Her menstrual cycle is regular. Her first menses occured at the age of 12. No vaginal discharge or blood clots reported. Multiple failed attempts of getting pregnant reported. Her last pap smear was normal 6 months ago. She reports weight gain of 10 pounds recently with an increased appetite. No cold intolerance, tremors, low mood, constipation, or bloating reported. She reports being anxious at work.

Example of Annotation in Patient Notes: Patient Notes highlighted with Annotations

Acknowledgement

The text data is provided by the National Board of Medical Examiners (NBME) from the USMLE Step 2 Clinical Skills examination, a medical licensure exam.