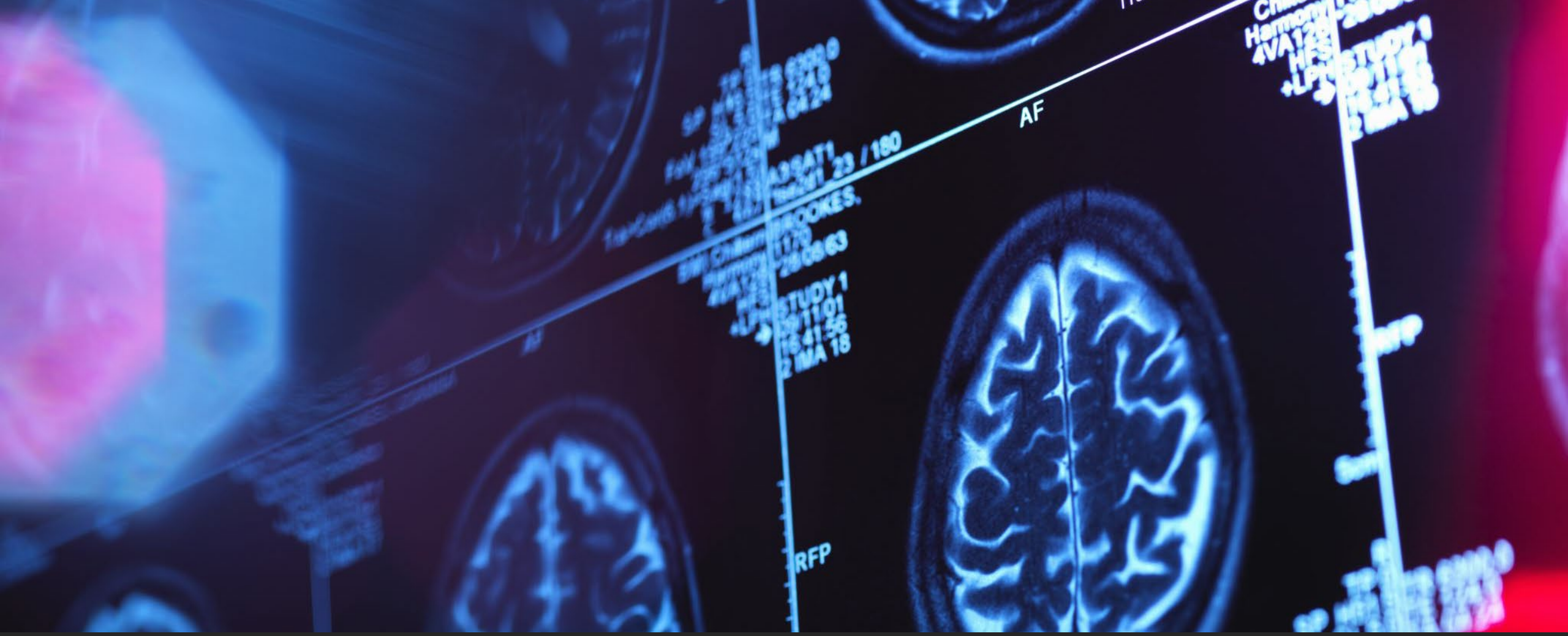


NLP for Healthcare – Classify Clinical notes

CSE6250 TEAM 75 – ANASTASIOS STATHOPOULOS,

SAGAR JOUNKANI,

ASHWIN MALGAONKAR,

LUIS MANUEL GARCIA BAQUERO CORREDERA

AGENDA

- Problem Statement
- Experimental Setup
- Data Preprocessing
- Data Exploration and Challenges
- Model Establishment
- Model Approach
- Model Selection
- Results
- Future Work

Problem Statement

Clinical notes are created all the time for every patient by clinicians and they include valuable information like diagnosis and medication treatment. They are always annotated by medical codes indicating the diagnosis and treatment applied. This process takes a lot of labor time since it requires analytical reading of the document from the medical staff and at the same time very good knowledge of the vast number of ICD codes.

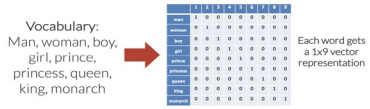
Every document may be tagged with many medical ICD codes increasing the time it takes to be annotated properly. In order to reduce this time a new approach will be implemented in this project as introduced in the paper (James Mullenbach, 2018) which fully automates the labelling process of a document of clinical notes.

The expected result would be the dramatical reduction of the time the labelling of such a document takes leaving for the clinicians only the validating process and at the same time the increase of reliability and accuracy of labelling.

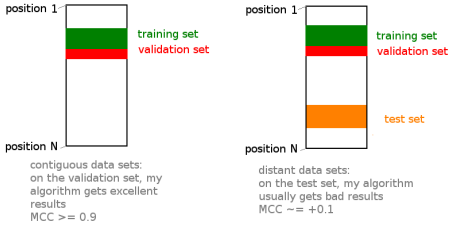
Experimental Setup

MIMIIC – 3 database

Used Spark scala and sql to preprocess clinical notes ,diagnoses and procedures data on local individual machines



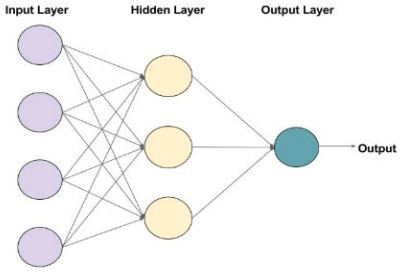
Word Embeddings



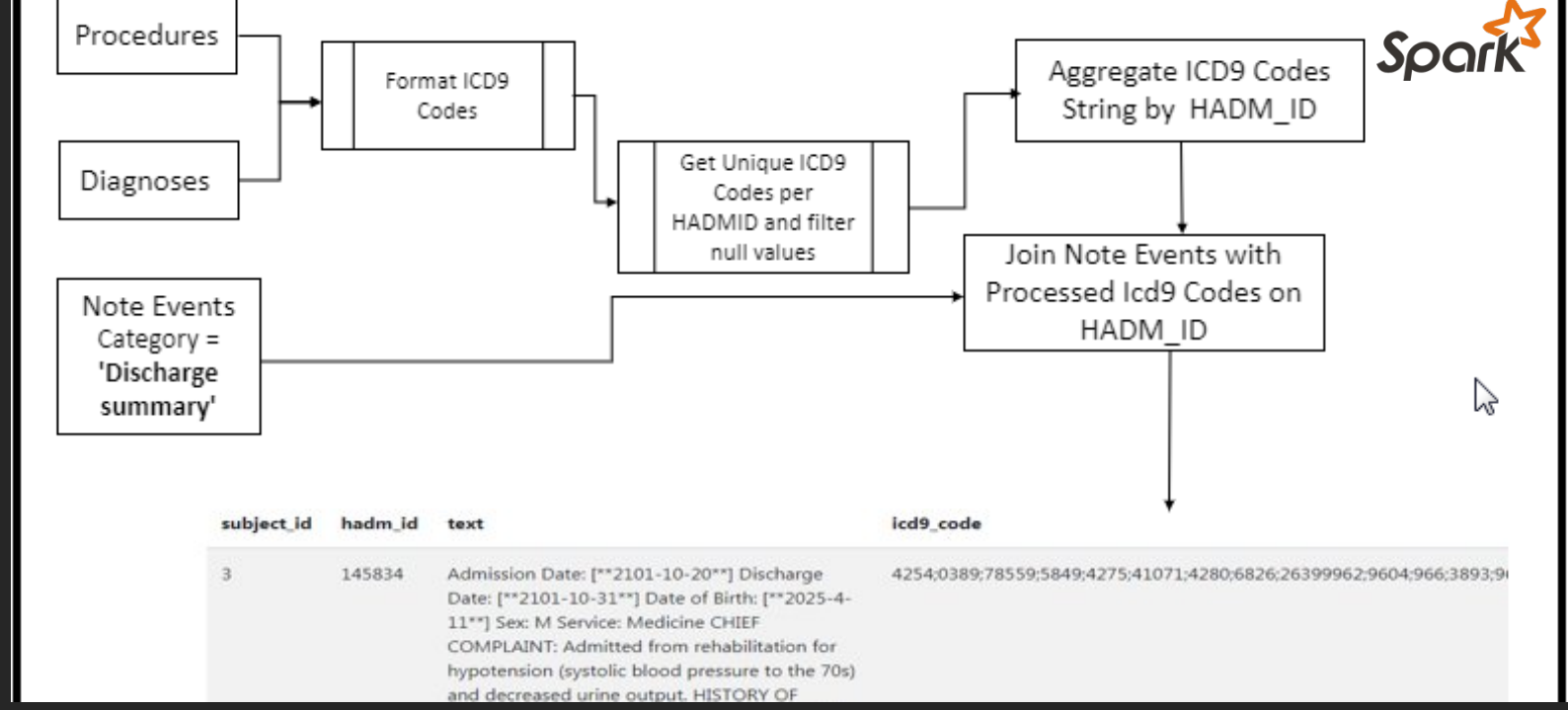
Training , Validation and Test Partition csv files

Google Cloud Platform

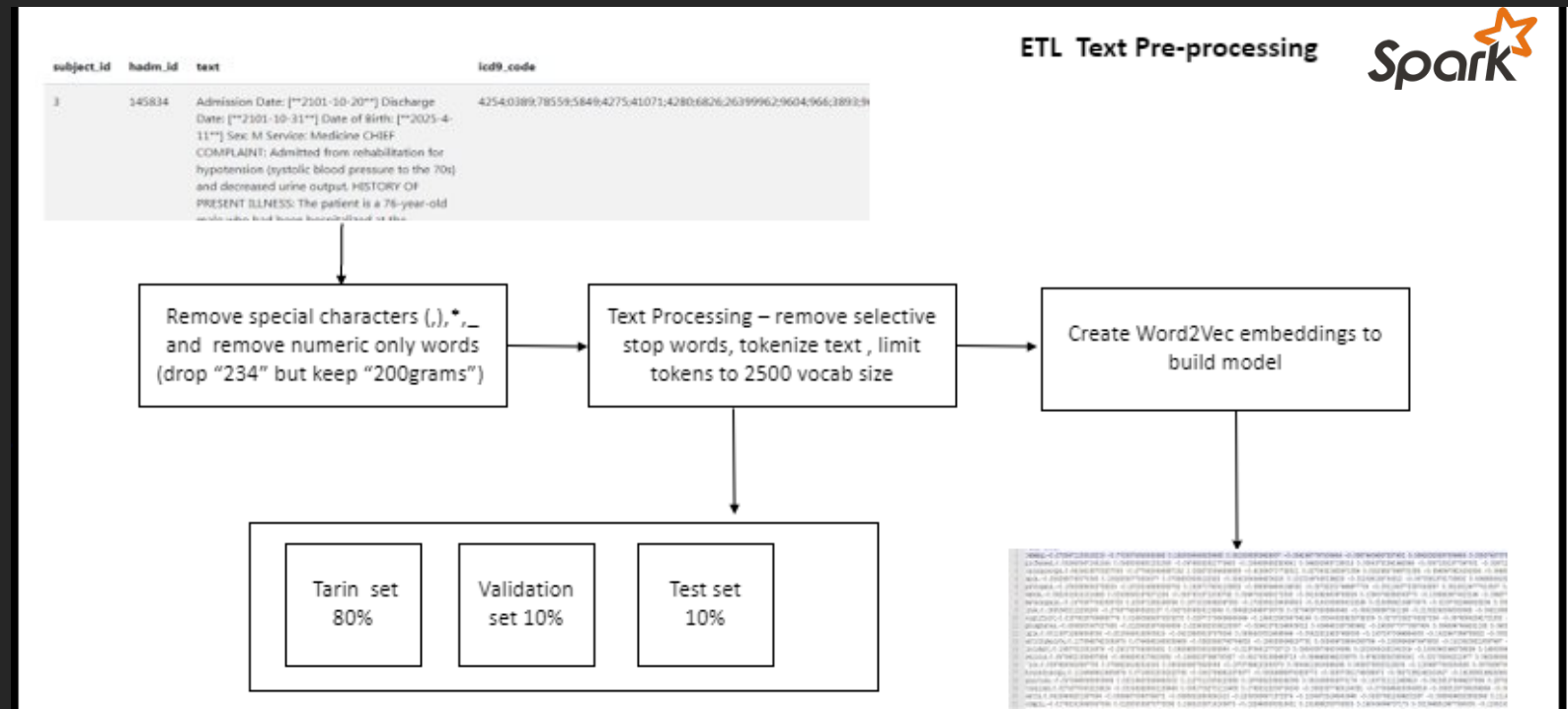
PyTorch



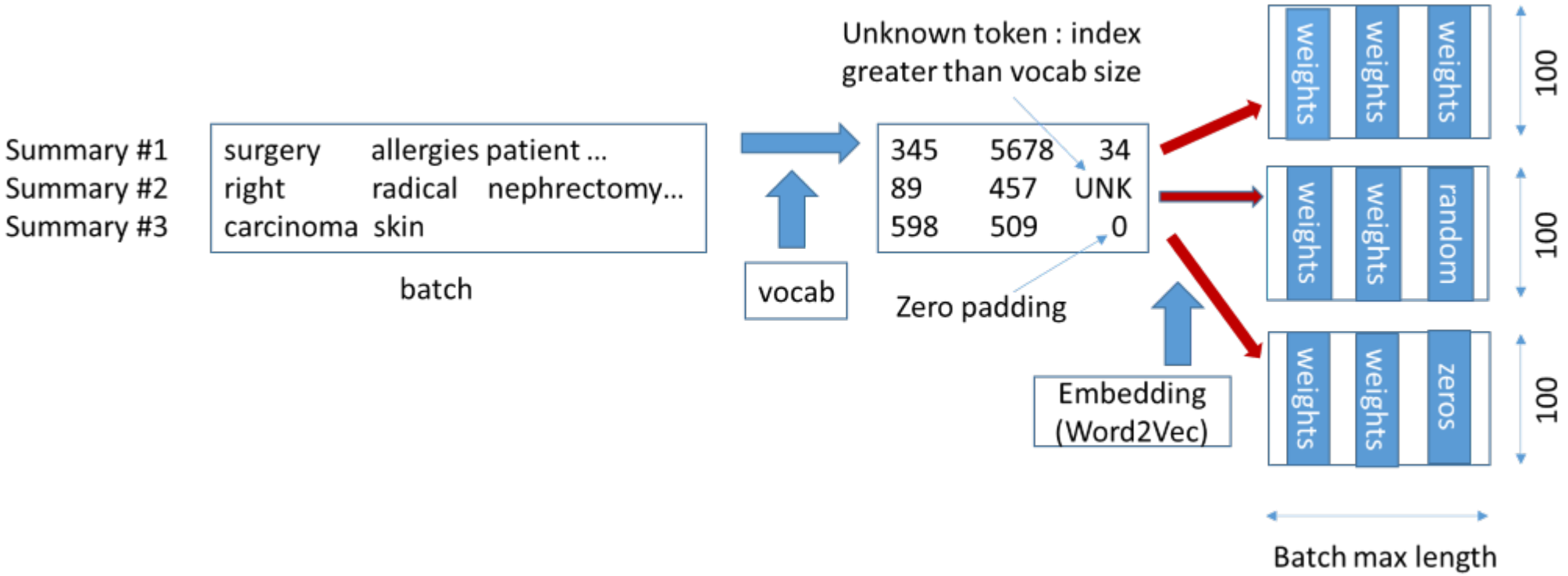
Data Preprocessing



subject_id	hadm_id	text	icd9_code
3	145834	Admission Date: [**2101-10-20**] Discharge Date: [**2101-10-31**] Date of Birth: [**2025-4-11**] Sex: M Service: Medicine CHIEF COMPLAINT: Admitted from rehabilitation for hypotension (systolic blood pressure to the 70s) and decreased urine output. HISTORY OF	4254;0389;78559;5849;4275;41071;4280;6826;26399962;9604;966;3893;9

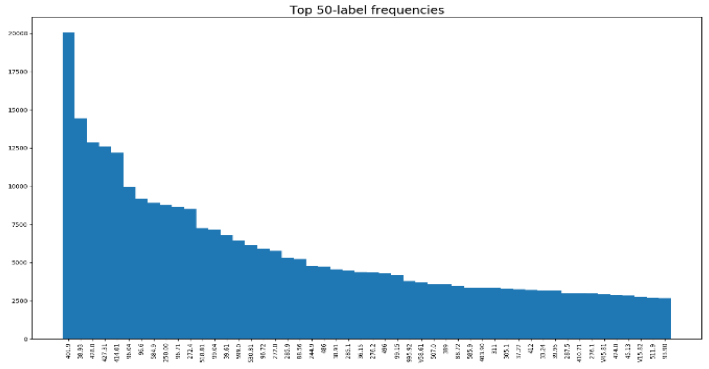


Feature Extraction



Data Exploration and challenges

- The loss function after 1st epoch dropped significantly and stayed the same throughout the training. Class imbalance was attributed to such behavior.



- Top-50 commonly used labels were sampled.

Model Establishment

- Binary Cross Entropy Loss With Logits

$$l_n = -w_n \left[y_n \cdot \log \sigma(x_n) + (1 - y_n) \cdot \log(1 - \sigma(x_n)) \right]$$

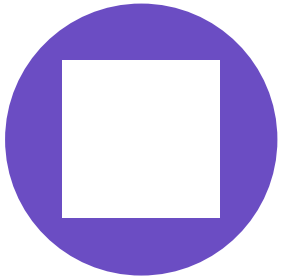
$$L = \{l_1, \ldots, l_N\}^\top$$

$$\ell(x, y) = \begin{cases} \text{mean}(L), & \text{if reduction} = \text{'mean'}; \\ \text{sum}(L), & \text{if reduction} = \text{'sum'}. \end{cases}$$

- Early Stopping included (with patience = 3)
- Probability Threshold : 0.5
- **Model Optimizer :** Adam Optimizer
- **Metrics used for evaluation :** Micro / Macro F1 Score , Accuracy , Recall, Precision and AUC

Model approaches

Logistic Regression : **The LR model is a multiclass classification network. It takes as input a tokenized set of words. Each of these tokens are mapped to a N dimensional latent space vector and stacked together. A max-pooling operation is applied on this matrix and pass it through a FCN layer with sigmoid activation and outputs as the number of classes.**



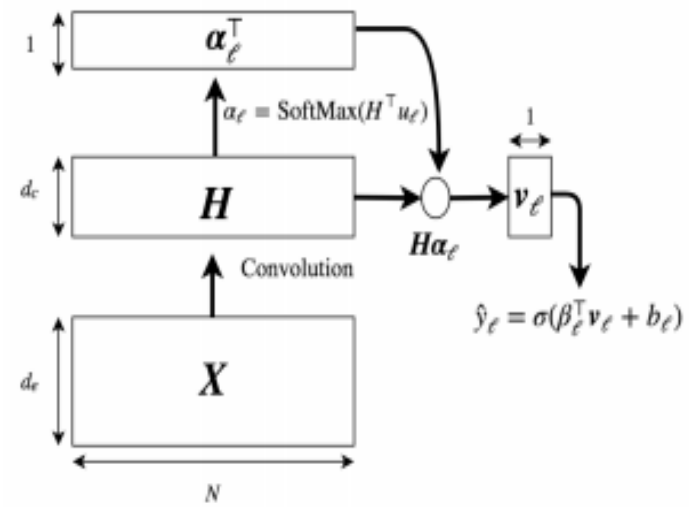
Vanilla CNN : The central intuition about this idea is to see hospital notes as images. For images, we also have a matrix where individual elements are pixel values. Instead of image pixels, the input to the tasks is sentences or documents represented as a matrix. Each row of the matrix corresponds to one-word vector.

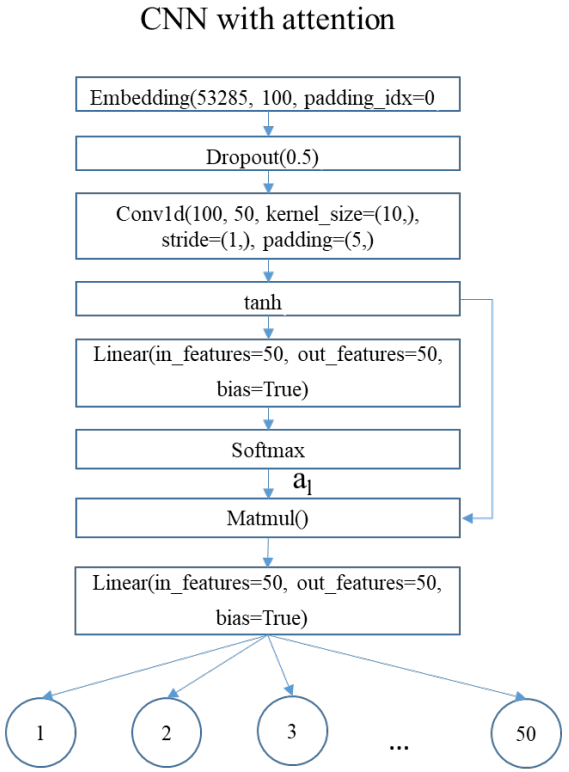
Bidirectional GRU : RNN learns words in close range and **understands** sequential context. They can remember previous information using hidden states and connect it to the current task. Bidirectional GRU keeps the contextual information in both directions which is **very** useful in text classification task.

CNN with attention mechanism : **Important** information is hidden in small snippets of text and for every label they are different**. We give different weights to different words for every label.**

Model Selection: CNN with attention mechanism

- Attention puts *the same sentence* along the columns and the rows, in order to understand how some parts of that sentence relate to others within text snippets in our document.
- Evaluated the tagging of the top 50 codes on sequences of different length and with a max length of 2500.
- Achieved macro precision of 0.642 and micro precision of 0.688.
- Important information is hidden in small snippets of text and for every label they are different.

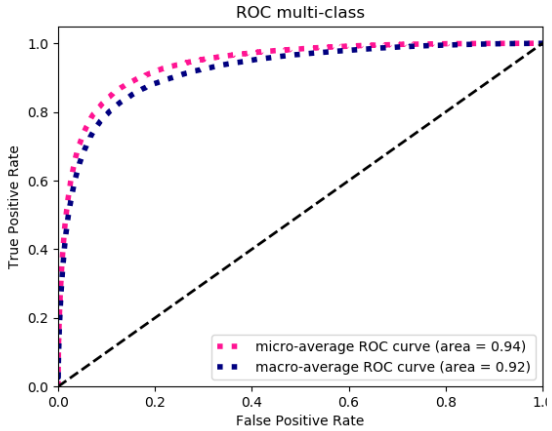




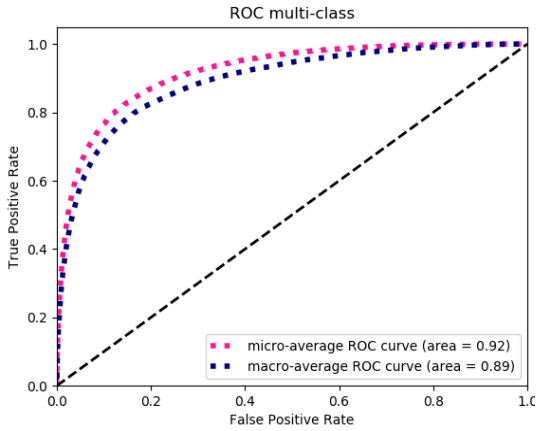
Results

Model	Macro Accuracy	Macro Recall	Macro Precision	Macro F1	Macro AUC	Micro Accuracy	Micro Recall	Micro Precision	Micro F1	Micro AUC
lr	0.288	0.349	0.632	0.450	0.888	0.375	0.446	**0.703**	0.546	0.914
Vanilla CNN	0.463	0.621	0.624	0.622	0.911	0.515	0.684	0.676	0.680	0.935
CNN attn	**0.483**	**0.637**	**0.642**	**0.640**	**0.921**	**0.526**	**0.690**	0.688	**0.689**	**0.941**
Bi-DI GRU	0.390	0.481	0.628	0.545	0.892	0.447	0.555	0.697	0.618	0.919

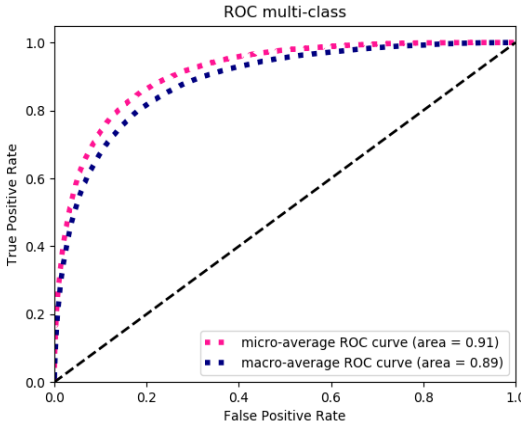
Results : ROC Curves



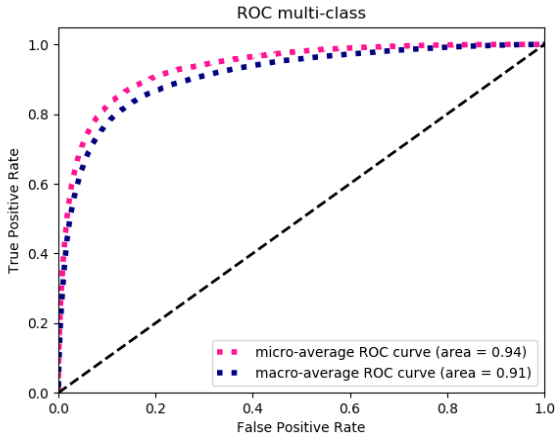
CNN with Attention



Bi-Directional GRU



Logistic Regression



Vanilla CNN

Future Work

- Introduce label description embeddings. Use the label descriptions and a regularization term in the loss function so for every label that is rare among the dataset, its parameters will be similar to those with similar description. This is a measure for imbalanced dataset especially for the full data set of more than 8000 unique ICD-9 codes
- Use word embeddings word2vec specially trained for medical/diagnosis sector
- Fine tune more the model
- Create RNN with attention mechanism to compare results

Thank you