Classification of Cancer Progression by Structuring Electronic Health Records

Chennuri Prateek | Chitipolu Gowtham | Girish Chandar G | Rahul Challa

Goals and Motivation

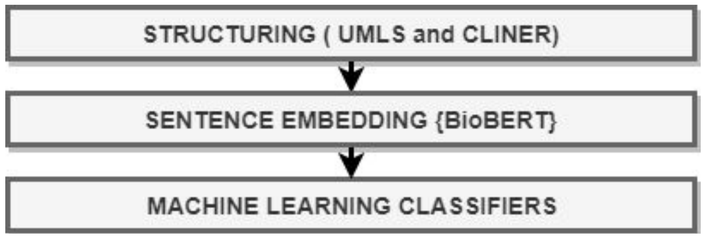
- Delay in diagnosis is one of the major causes of cancer related deaths.
- Electronic Health Reports (EHRs) have more information than radiological reports and are also available for every patient visiting the hospital unlike radiological report.
- We aim to apply NLP techniques to structure the EHRs and then predict the cancer probability.

Dataset

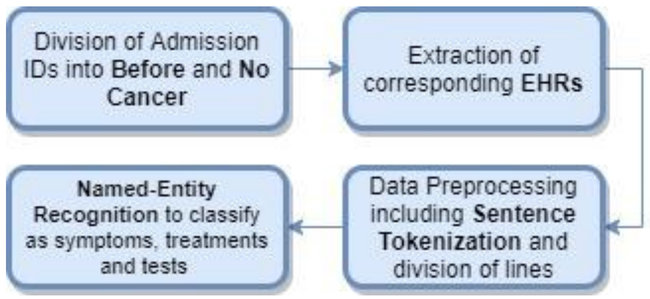
- Dataset used: **MIMIC-III (Medical Information Mart for Intensive Care - III)**[1].
- It contains de-identified data of ten years of around 40,000 patients in which 7,361 patients are diagnosed with cancer.
- Each patient has multiple EHRs, which are in the form of unstructured data.
- Each patient may have multiple Admission IDs depending on their medical history.

Total number of patients	46520
Total number of cancer patients	7361
Total number of discharge summaries	59456
Discharge summaries of cancer patients	11495
Discharge summaries before the diagnosis of cancer	1949
Discharge summaries after the diagnosis of cancer	9546

Pipeline



Structuring



- Classifier training requires data samples with both the labels, **Before** and **No Cancer**.
- CliNER is a NLP processing system specifically designed for **Named-Entity Recognition** for clinical files.
- **CliNER**[4] and **UMLS**[3], tag specific words from the EHRs as problem, treatment or test.
- Extract corresponding lines from the EHRs containing the specific words for embedding.

Sentence Embedding

- We used **BioBERT**[2] to embed each sentence in the EHR.
- BioBERT embedding for an EHR is done separately for the three tags(problems, treatments, tests)
- The mean of all the sentence embeddings of an EHR is considered as the embedding which represents the whole EHR.

Machine Learning

- We used Deep Neural Network model which is trained on these embeddings as input and their classes (Before or No cancer) as labels.
- We have also tried using other classifiers like SVM, KNN and Gaussian NB.

Results and Discussions

Model	F1 Score	Accuracy (%)	Precision	Recall
Neural -1	0.59	65.5	0.5191	0.6834
Neural -2	0.663	63.36	0.63	0.701
Neural -3	0.6473	66.66	0.642	0.6527
Neural -4	0.5924	67.53	0.4945	0.738
Neural -5	0.614	65.7	0.578	0.657
Gaussian NB	**0.669**	58.27	0.8852	0.5382
Overall	0.655	**68.90**	0.6475	0.663

Based on the structuring elements used we have the following metrics (The architecture of the neural network remains constant)

Metric	Problems	Problems+ Treatments	Problems+ Treatments+ Tests
Accuracy	0.6327	**0.688**	0.638
Precision	0.5737	0.6475	0.297
Recall	0.688	0.663	0.844
F1 Score	0.6259	**0.655**	0.44

References

[1] Johnson, Alistair EW, et al. "MIMIC-III, a freely accessible critical care database." *Scientific data* 3 (2016): 160035.

[2] Lee, Jinhyuk, et al. "Biobert: pre-trained biomedical language representation model for biomedical text mining." *arXiv preprint arXiv:1901.08746* (2019).

[3] Bodenreider, Olivier. "The unified medical language system (UMLS): integrating biomedical terminology." *Nucleic acids research* 32.suppl_1 (2004): D267-D270.

[4] Boag, William, et al. "CliNER: a lightweight tool for clinical named entity recognition." *AMIA Joint Summits on Clinical Research Informatics (poster)* (2015).

Acknowledgements

We would like to thank Prof. Mayank Singh and our TA Pratik Kayal for their continued inputs and support during the course of the project.