CLIFT : Analysing Natural Distribution Shift on Question Answering Models in Clinical Domain

Ankit Pal
Research Engineer at Saama AI Research Lab
https://aadityaura.github.io

NEURAL INFORMATION PROCESSING SYSTEMS

Problem Statement

In statistical learning theory, it is often assumed that test data comes from the same distributional underpinnings as training data for a machine learning model. However, in reality, data may be sampled from various distributions with considerable domain changes, i.e. Federated learning [1]. When data distribution unexpectedly changes, models that seem to be doing well in test set accuracy, fail when deployed in real time. Since healthcare data is very sensitive, Real-world deployments of health machine learning systems may encounter these failure situations. Robust machine learning aims to create techniques that consistently work in different environments.

$$l_Q(M) - l_{Q'}(M) = \underbrace{(l_Q(M) - l_{\mathcal{P}}(M))}_{\text{Adaptivity gap}} + \underbrace{(l_{\mathcal{P}}(M) - l_{\mathcal{P}'}(M))}_{\text{Distribution gap}} + \underbrace{(l_{\mathcal{P}'}(M) - l_{Q'}(M))}_{\text{Generalization gap}} \quad (1)$$

Where M is the model, l is a loss function, Q is the test dataset, p is the data distribution and P', Q' are the additional data distribution and test set respectively.

Proposed Testbed

In this paper, we propose a testbed called CLIFT (**Cli**nical Shi**ft**) to evaluate the clinical deep learning model's performance under the distributional shift of the different test sets. It will help to make the clinical ML models more robust. In brief, the contributions of this study are as follows.

- We are proposing five new QA test datasets, including Cancer, Heart, Smoke, Obesity & Medication
- Covering different clinical domain diseases and multiple sub-topics
- These samples are from MIMIC-III clinical database
- Detailed statistics and evaluation of natural distribution shift on emrQA
- Open-source code and model checkpoints to reproduce the results

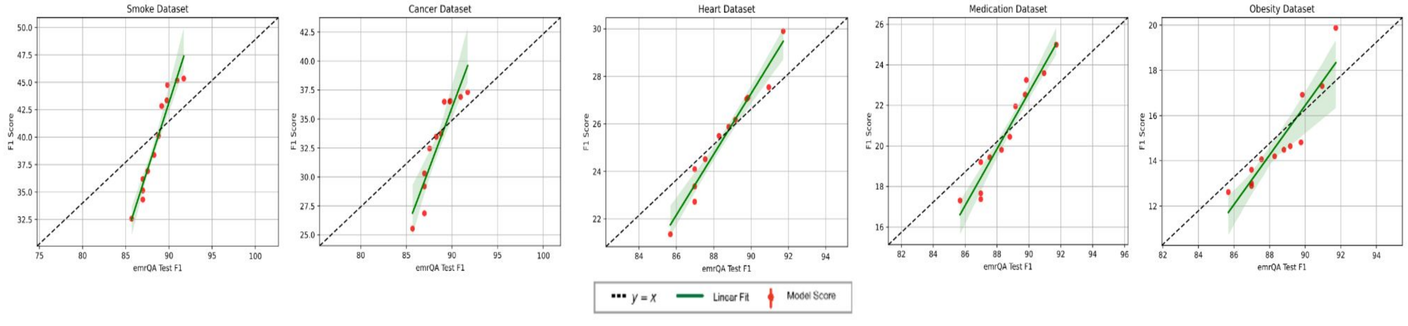

Figure 2 : Comparison of models' F1 scores on emrQA and our proposed test datasets.

Dataset Statistics & Analysis

\

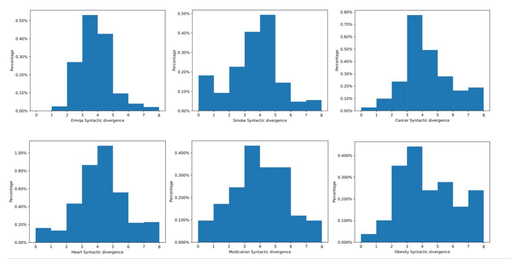

Figure 1: Top row shows the histograms of the emrQA, Smoke and Cancer testset. The bottom row shows the Heart, Medication and Obesity histograms.

	Smoke	Heart	Medication	Obesity	Cancer
Question #	2k	1.5k	1.5k	1k	1.5k
Min Q tokens	2	3	3	3	3
Min A tokens	1	2	1	1	1
Min C tokens	100	200	100	100	100
Avg Q tokens	6.42	8.31	7.61	7.19	8.40
Avg A tokens	4.59	4.15	3.93	4.04	4.12
Avg C tokens	217.33	234.18	215.49	212.88	210.16

Table 1 : statistics of the dataset

Table 1 summarizes the general statistics of preprocessed data. We plan to expand the corpus by adding more samples and model results. The full paper and the updated benchmark are available at **https://openlifescience-ai.github.io/clift**

Experiment Results

Model	emrQA F1	Clift F1	Gap (δ)
BioBert-B	87.55	32.57	54.98
BERT-ClinicalQA	88.28	34.31	53.97
Electra-Squad-L	86.98	35.13	51.85
DistilBert-squad-B	85.68	36.18	49.51
BioClinicalBert	86.98	36.90	50.08
Biomed-Roberta	89.76	38.38	51.39
BlueBert-PubMed	86.98	40.11	46.87
Bert-Large	88.81	42.84	45.97
BioBert-squad-L	90.95	43.36	47.60
PubMedBERT	91.72	44.75	46.97
Roberta-squad-B	89.84	45.16	44.68
BlueBert-MIMIC	89.16	45.34	43.82

Table 2 : Performance of different models on the emrQA Testset and Smoke Testset

We compared the F1 scores for all the models trained on the emrQA to the F1 scores on each of our additional test sets. Figure 2 shows that, when applied to proposed test datasets, All the models indicate a performance loss ranging from 43.82% to 73.07%. Table 2 shows the performance of different models on the Smoke test set. As we can see that BioBert-B drops around 54% F1 points while BlueBert-MIMIC drops around 43%

These results imply the potential for further study handling distribution shifts in clinical datasets. In addition, future research should look at specific model training parameters and the datasets' size, quality, and quantity that contribute to these variations. Our extensive testbed will indicate progress towards trustworthy machine learning in the healthcare area.

References

[1] Arthur Jochems (2016). "Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital - A real life proof of concept." In: Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology

[2] Anusri Pampari (2018). "emrQA: A Large Corpus for Question Answering on Electronic Medical Records". In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing