BMI 598 Final report
Task 7 - Natural Language Inference in medical domain

Task Review

The main task of our project encompassed several subtasks using the provided MedNLI dataset such as data preprocessing and ultimately comparing the results of our approach to the baseline results (also current SOTA) and analyzing results on the MedNLI test set. Table 1 shows the examples for each of the class. Where E is entailment, C is contradiction and N is neutral.

Premise	Hypothesis	Class
No history of blood clots or DVTs, has never had chest pain one week ago.	Person A has angina.	E
Over the past week PTA he has been more somnolent and difficult to arrouse	Over the past week he has been alert and oriented	C
During hospitalization, patient became progressively more dyspnic requiring BiPAP and then a NRB.	The patient has pulmonary edema.	N

Table 1: Examples for each class

Method

Proposed method

In our proposed method we used a variant of base uncased BERT (BlueBert) that is pretrained on the corpus of *MIMIC* and *pubmed* abstracts. Our main goal was to use BlueBert for our NLI task and compare the results with baseline methods and current State of the art. Furthermore, we used an additional fine-tuning methodology where the Med-NLI dataset was transformed to include medical abbreviation information using the abbreviations dataset from [4].

Innovation

To improve the performance of our model we used the multi-task learning approach by using all three datasets (SNLI, MNLI, Med-NLI). To achieve we iteratively fine tune BlueBert on mixed samples from the SNLI and MNLI datasets respectively and later on on the MedNLI dataset individually. Additionally, we also implemented medical abbreviations data infusion to record any possible improvements in the performance. As part of our post prediction error analysis, we also tried to interpret the model findings and errors. To further improve the model performance, we tried abbreviation data infusion using 2 different approaches to infuse abbreviation meanings. In the first approach the abbreviations were explicitly mentioned as terms so that the attention heads can learn the co-references better and in the second approach the abbreviations were replaced by their meaning directly. However, we achieved better results with the second approach.

Experiment setup

Baseline and current SOTA Method

The baseline method discussed in Alexey et al. [1] shows accuracies of different models (CBOW, Bi-LSTM, ESIM) on NLI corpora (MultiNLI and SNLI). In this project, we compared the performances of pre-trained CBOW and Bi-LSTM on MedNLI dataset with our main model's performance on the same. The current State of the Art architecture and methodology is described by Boukkouri et al. [5] where they implement a BERT architecture on a character level using a Character-CNN module[6]. Character level Bert provides more robustness in case of noise and misspellings.

Data Separation

For our NLI task we use MultiNLI, SNLI and the MedNLI datasets. The training data contains around 11000 records and the validation files contains around 2000 records. The test set contains around 1400 samples. To enable our multi-task learning approach, we use around 250000 samples from the mixed set of SNLI and MultiNLI respectively for the training phase and around 20000 samples from the validation set of MNLI for the development set. As per the instructions, we use the mentioned test set of around 1422 samples from the MedNLI dataset as our test set.

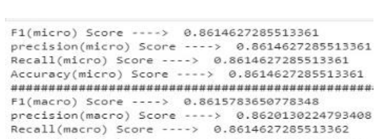
Evaluation Metric

We use **Accuracy** and the **F-1 measure.** As, the class distribution in our test set is uniformly balanced, we intend to use accuracy as one of our classification metrics.

Results

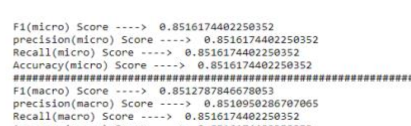
Results on the test set

Figure 1(a) and Figure 1(b) show F1 scores for the initial and improved approach respectively. Additionally, Table 1 provides a comparative performance comparison between various Baseline Methods (and current SOTA) and BlueBert. As, it can be seen, multi-task Learning from SNLI and MNLI helps achieve State of the art results.



```
F1(micro) Score ---->  0.8614627285513361
precision(micro) Score ---->  0.8614627285513361
Recall(micro) Score ---->  0.8614627285513361
Accuracy(micro) Score ---->  0.8614627285513361
##################################################
F1(macro) Score ---->  0.8615783650778348
precision(macro) Score ---->  0.8620130224793408
Recall(macro) Score ---->  0.8614627285513362
Accuracy(macro) Score ---->  0.8614627285513361
```

Figure 1(a): Evaluation metrics for Multi-task learning model



```
F1(micro) Score ---->  0.8516174402250352
precision(micro) Score ---->  0.8516174402250352
Recall(micro) Score ---->  0.8516174402250352
Accuracy(micro) Score ---->  0.8516174402250352
############################################################
F1(macro) Score ---->  0.8512787846678053
precision(macro) Score ---->  0.8510950286707065
Recall(macro) Score ---->  0.8516174402250352
Accuracy(macro) Score ---->  0.8516174402250352
```

Figure 1(b): Performance metric for Abbreviation Infused Mult learning model

Error Analysis

Figure (2) and Figure (3) show the confusion matrix and the error plot for the test set respectively. It would definitely be more encouraging to get prediction error in the midway (0.3-0.7) ballpark.

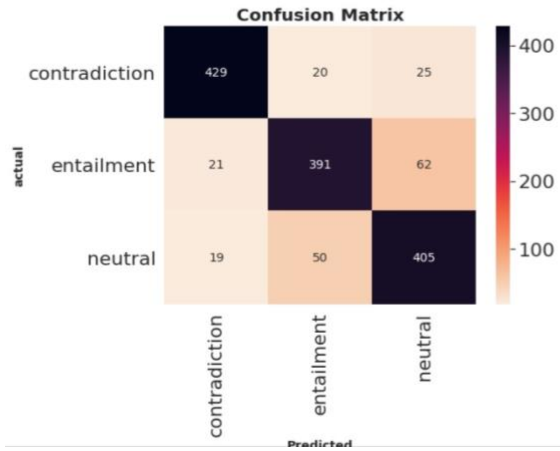


Figure 2: Confusion matrix

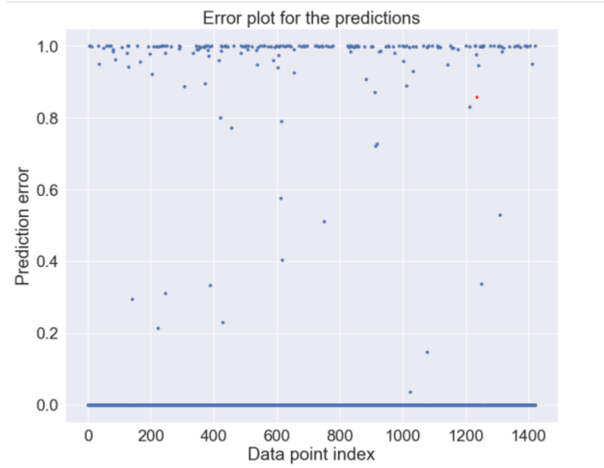


Figure 3: Error plot for all data points

Table 3 depicts certain categories of errors where the model fails at. All of the premise hypothesis pairs are categorized into one of the three categories shown in Table 4.

Category	Premise	Hypothesis	Explanation
1	HISTORY OF PRESENT ILLNESS: This is an 80-year-old male with old ischemic infarction and some residual right hemiparesis, hypertension, and atrial fibrillation (on warfarin with an INR of 2.5) who presented with increasing right-sided weakness on [**3193-2-22**].	History of hypertension	The model fails to apply temporal reference in this case the word "history" adds a temporal reference to the hypertension.
1	Nutmeg liver reaction, common bile duct within normal limits.	patient has normal liver	Normal limit refers to bile duct and not liver, however, the model couldn't understand it.
1	One 15 year old son helps her, and apparently this is a problem for the patient, either because the patient is concerned about his back, or because she does not always get the help she needs, this is unclear.	The patient needs help with transfers.	The patient might be needing help but not with transfers. However, it is important to have a reference of help with transfers and this needs to be learned as help and transfers are both good keywords
2	Congestive heart failure.	Patient has abnormal ejection fracture	Lack of general medical knowledge resulted in ejection fraction not being associated with heart failure.
2	On field, pupils were sluggish 2-3 mm but responsive, HR 100, RR 4, no audible BP, pt pale, cool and diaphoretic.	the patient has a low respiratory rate	Lack of general medical knowledge.
2	Status post dilatation and curettage.	Patient was recently pregnent	Association of pregnancy with dilation and currettage lacking and also not robust to misspellings
3	During that study, the patient developed acute pulmonary edema which required treatment with Morphine and Lasix.	She developed shortness of breath	Using the co-occurences to infer premise to hypothesis, meaning as pulmonary edema is associated with shortness of breath, it gave entailment but it necessarily doesn't have to be related.
3	She was eventually found to have multiple problems including pericardial effusion, cardiac tamponade, a small pleural effusion, swollen knee with a joint effusion and joint pain.	She has arthritis	Using the co-occurences to infer premise to hypothesis - joint effusion and swollen knee is being associated with arthritis. This may not be true if other symptoms prevail.
3	The history is per the pt and her granddaughter who was present for the events.	The patient did not give the history.	Inferencing power: As the history has been gotten from PT and granddaughter, it takes the inferenceing power to understand that the patient did not give the history.

Table 3: Error categories and examples

Error Category Number	Error Category
1	Object referencing/Temporal referencing
2	Lack of medical general knowledge
3	Inference reasoning on basis of co-occurences

Table 4: Error Categories Description

Conclusion:

We performed multi-task learning with the pre-trained BlueBERT model along with medical abbreviation data infusion. Our vanilla multi-task learning version of BlueBERT has produced State of the Art results with the F1 Score and Accuracy of 86.16% and 86.15% respectively. As we can see that even though the medical data infusion did not perform as expected, an expansive abbreviation dataset along with more training samples can definitely help address issues mentioned in our error categories.

References

[1] "Lessons from Natural Language Inference in the Clinical Domain" Alaxey Romanov et al. https://arxiv.org/pdf/1808.06752.pdf

[2] "Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets" Yifan Peng et al., https://arxiv.org/abs/1906.05474

[3] Baseline Models for MultiNLI Corpus, https://github.com/nyu-mll/multiNLI

[4] Medical abbreviations dataset, https://www.kaggle.com/eeemonts/medical-abbreviations

[5] CharacterBert: Current State of the Art, https://arxiv.org/pdf/2010.10392v3.pdf

[6] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 2227–2237, New Orleans, Louisiana, June. Association for Computational Linguistics.