Assertion Status Classification Using BiLSTM-CRF

With and without information on section types

Kexin Chen, Mar 13, 2024

Previous study

The bars in the graph show the distribution of assertion classes per section type. Concepts in these four sections are more likely to be "absent" compared to concepts in other sections: *Physical examination/Status*, *RoS*, *Allergies*, *Complications*. More than 50% concepts in these sections are absent.

(Note: Section headings are extracted using regular expressions.)

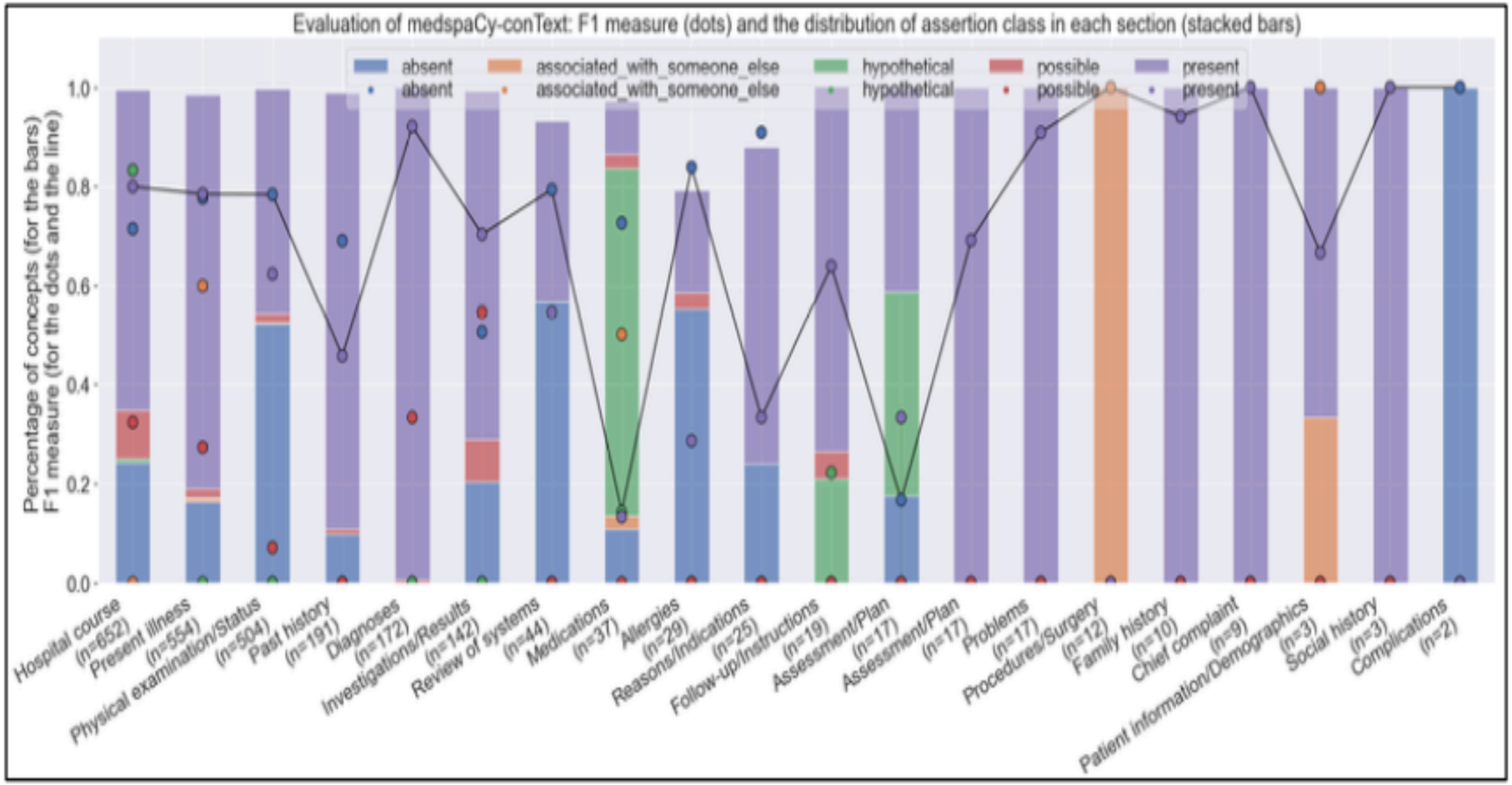


Figure 1. Stacked bars: Distribution of assertion classes per section. They do not add up to 1 because model's unability to determine "conditional" concepts. Dots: F1 measure of each assertion class. The black line connects the dots that represent the most frequently occurring assertion class in each section. This line serves as a reference point for comparison. Colors: Assertion classes.

2010 i2b2 Challenge

- Task: Assigning assertion types for medical problem concepts.
- Note type: Discharge summaries and progress notes
- A total of 826 clinical notes are split into 280, 69, and 477 for train, dev, and test, respectively.
- The dataset is annotated with medical problem concepts and their assertion status, i.e. present, absent, possible, conditional, hypothetical, and associated with someone else.

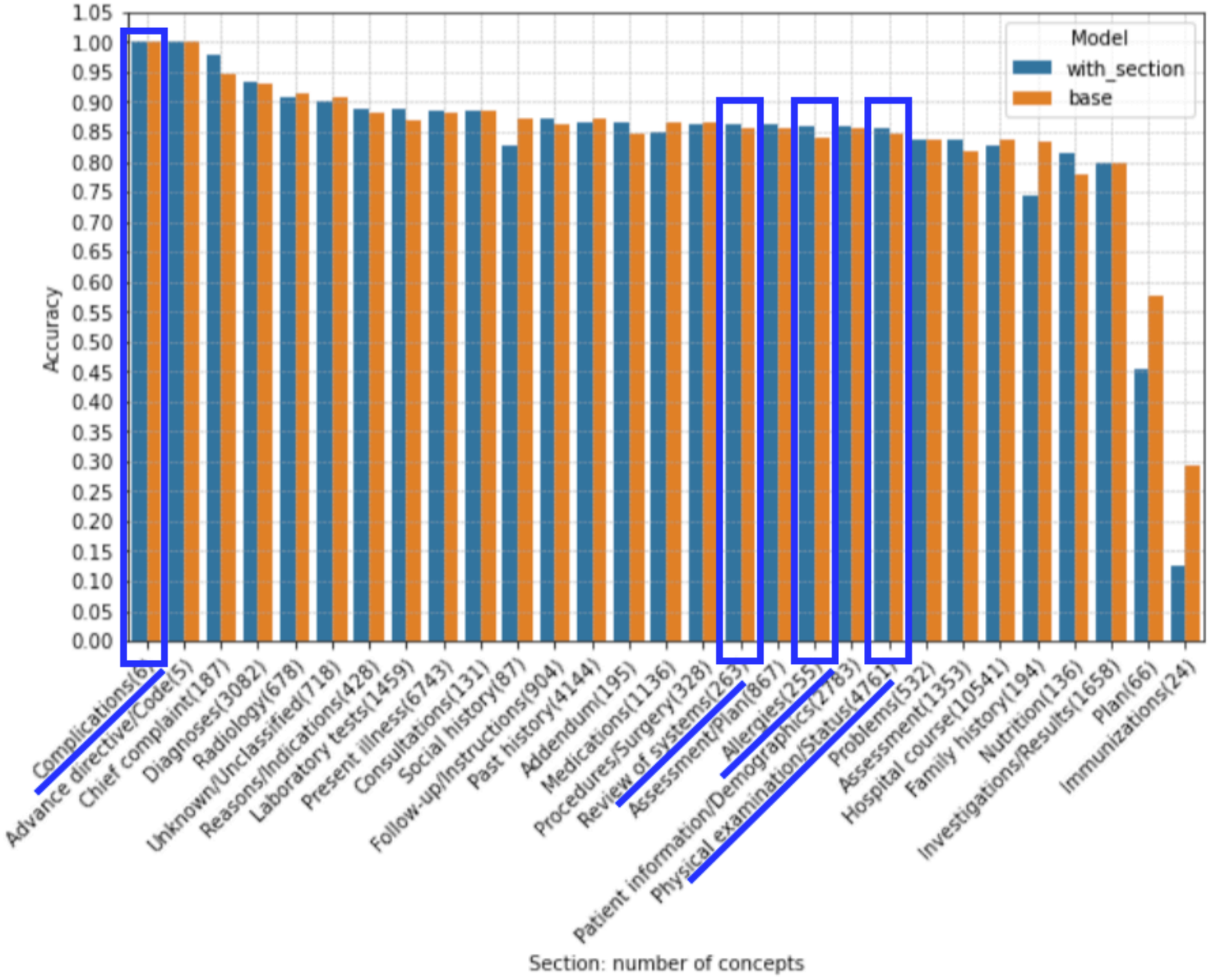
BiLSTM-CRF base model vs using section type as additional input

- Two models are trained:
  - BiLSTM-CRF (base)
  - BiLSTM-CRF using section type as additional input
- Post-processing:
  - Classify multi-token concepts based on the majority token category.
  - Relabel "N/A" concepts as “present.”
- Using section info in the model seems to reduce labeling concepts as N/A.

Table 1

	Evaluated on ... (# of concepts)	Accuracy (Relabeled)	Accuracy	Labeled as N/A	# of Concepts
BiLSTM-CRF Base	Full	0.9284	0.8708	0.0700	18550
Use section type	Full	0.9261	0.8708	0.0666	18550
BiLSTM-CRF Base	Concepts in the highly-negated sections	0.9216	0.8656	0.0817	2411
Use section type	...highly-negated sections	0.9270	0.8781	0.0693	2411
BiLSTM-CRF Base	Concepts in the lowly-negated sections	0.9294	0.8716	0.0682	16139
Use section type	... lowly-negated sections	0.9260	0.8698	0.0662	16139

- The graphs shows the two models' performance on concepts in different sections.
- It is not obvious whether using section information will increase accuracy for concepts in the lowly-negated (<40%) sections.
- Using section information slightly helps classifying the assertion status of concepts in the highly-negated sections.



Models trained on the full corpus vs trained on a subset (i.e. sentences in the highly/lowly-negated sections)

- Sentences are tagged as “highly-negated” if they belong to one of the four highly-negated sections. They form the “highly-negated” dataset. All other sentences form the “lowly-negated” dataset.
- BiLSTM-CRF models are trained on each subset with and without using section information as an additional input. Results are shown in Table 2.
- Model trained on the smaller subset (i.e. the highly-negated subset) shows lower accuracy.
- Using section info in the model seems to reduce labeling concepts as N/A.

Dataset	Description	Train	Dev	Test
Full	The original set	236,113 tokens	56,386 tokens	29,744 sents 441,950 tks
Highly-negated	Sentences that belong to the highly-negated sections	1,553 sents 21,919 tokens	286 sents 4,037 tokens	2,715 sents 36,316 tokens
Lowly-negated	Sentences that belong to all other sections	14,126 sents 214,192 tokens	3,518 sents 52,347 tokens	27,029 sents 405,629 tks

Table 2

Method	Trained on which dataset (or subset)	Accuracy (Re-labeled. On the corresponding test set)	Accuracy (On the correspondi ng test set)	Labeled as N/A	# of Concepts
BiLSTM-CRF	Full	0.9284	0.8708	0.0700	18550
BiLSTM-CRF + section type	Full	0.9261	0.8708	0.0666	18550
BiLSTM-CRF	Concepts in the highly-negated sections	0.8880	0.8287	0.0892	2411
BiLSTM-CRF + section type	...highly-negated sections	0.8864	0.8308	0.0888	2411
BiLSTM-CRF	Concepts in the lowly-negated sections	0.9163	0.8737	0.0509	16139
BiLSTM-CRF + section type	... lowly-negated sections	0.9239	0.8532	0.0858	16139
Combine		0.9078	0.8635	0.0551	18550
Combine		0.9146	0.8465	0.0853	18550

Down-sample the datasets and train models on the smaller subsets for fair comparison - Dataset

- ~10% sentences are randomly selected to form the “downsample-full” set.
- ~10% sentences from the lowly-negated sentences are randomly selected to form the “downsample-lowly-negated” dataset.
- In this way, the size of training data is the same across all three datasets (i.e. highly-negated, downsample-full, downsample-low-negated)

Dataset	Description	Train	Dev	Test
Full	The original set	236,113 tokens	56,386 tokens	29,744 sents 441,950 tks
Highly-negated	Physical examination/ Status, RoS, Allergies, Complications	1,553 sents 21,919 tokens	286 sents 4,037 tokens	2,715 sents 36,316 tokens
Lowly-negated	Sentences that belong to other sections	14,126 sents 214,192 tokens	3,518 sents 52,347 tokens	27,029 sents 405,629 tks
Downsample-Full	~10% of all the sentences	1,553 sents 23,293 tokens	286 sents 4,439tokens	2,715 sents 38,899 tokens
Downsample-lowly-negated	~9% of the lowly-negated sentences	1,553 sents 23,198 tokens	286 sents 4,471 tokens	2,715 sents 38,824 tokens

Down-sample the datasets and train models on the smaller subsets for fair comparison - Results

- Adding section type as an additional input does not have impact accuracy.
- Compared to models trained on the full size of data, models trained on the small subsets shows obviously worse performance. Apart from the impact of decreasing data size, another potential reason is that randomly selected sentences lack consistency in meaning. Sentences in the highly-negated sections remain relatively complete and consistent, and thus have better performance.

Table 2

Method	Trained on which dataset (or subset)	Accuracy (on the corres-ponding test set)	Accuracy (On the correspondi ng test set)	Labeled as N/A	# of Concepts
BiLSTM-CRF	Downsample full	0.8393	0.6791	0.2067	1630
BiLSTM-CRF + section type	Downsample full	0.7325	0.7092	0.1693	1630
BiLSTM-CRF	Highly-negated	0.8880	0.8287	0.0892	2411
BiLSTM-CRF + section type	Highly-negated	0.8864	0.8308	0.0888	2411
BiLSTM-CRF	Downsample lowly-negated	0.7552	0.5551	0.2589	1634
BiLSTM-CRF + section type	Downsample lowly-negated	0.8427	0.7607	0.1065	1634

Summary

- Using section information as an additional input when training models slightly helps classifying assertion status of the concepts in the *highly-negated* sections. But after converting N/A values, the impact of sections seems to be outweighed.
- Training different models on different sections leads to worse performance, possibly because disadvantages of training on a smaller dataset (especially, the highly-negated subset) outweighs the benefit of developing specific models for different sections. My experiment design cannot lead to conclusions that when training data is sufficient, whether training different models on different sections improves accuracy or not.

Downsample

Hide concepts in the non-target sections

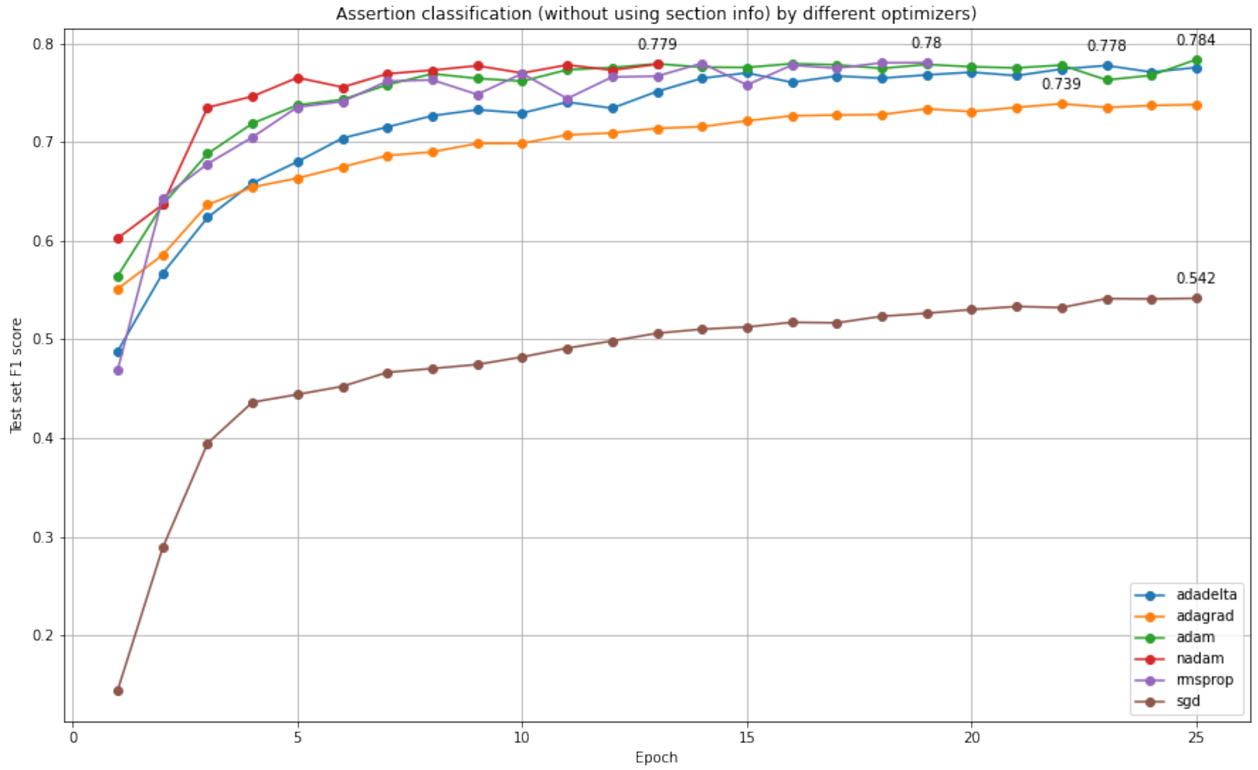
Appendix: Previous reports

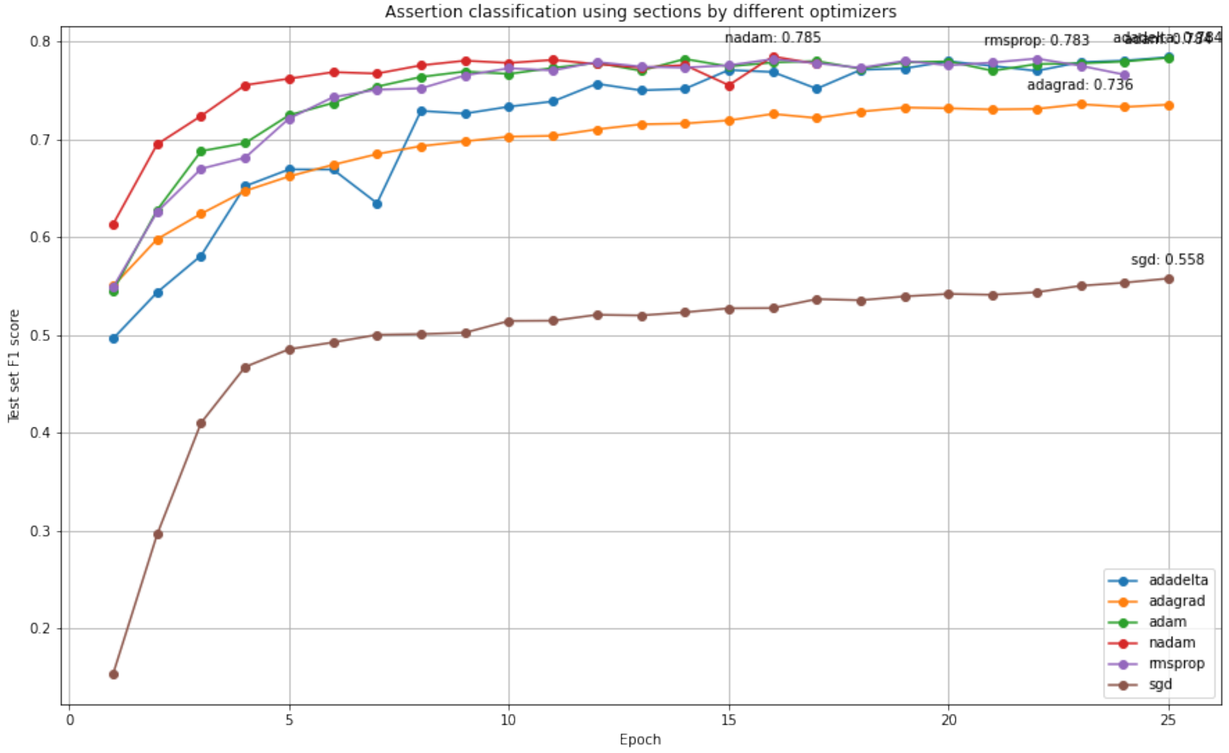
(Wrong) Task: Identify concepts and its negation status
Develop a model that identifies “problem” concepts AND their contextual attributes.

(Wrong) Task: Identify concepts and its negation status

Different optimizers

- Conclusion: the nadam optimizer is quick and well-performing compared to adadelta, adagrad, Adam, rmsprop, and sgd.

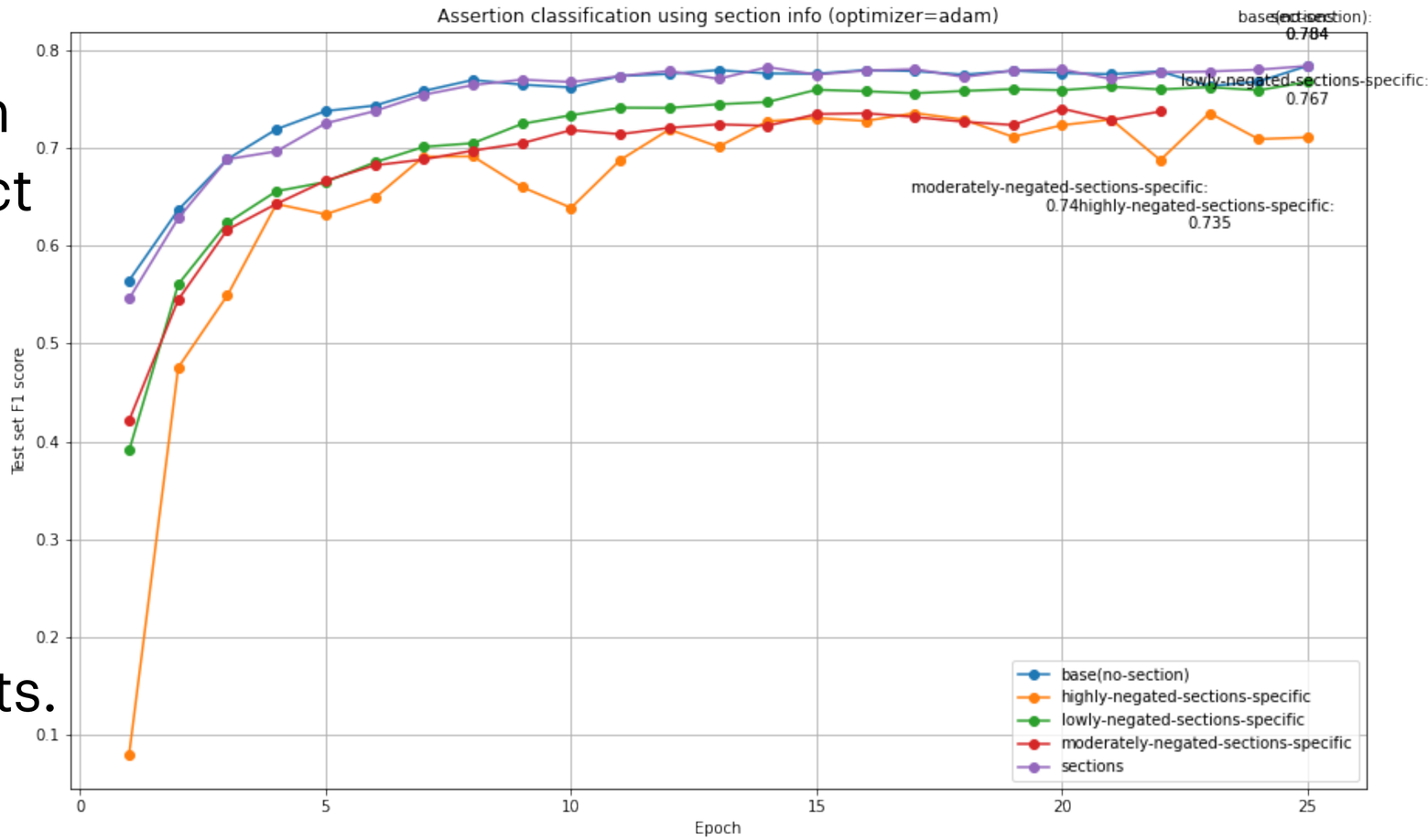




(Wrong) Task: Identify concepts and its negation status

Explore 1. Train the same model on sections that contain more prevalent negated concepts
Explore 2. Include section type of each token as an input column .

- Whether to Include section type or not does not impact performance.
- Model has worse performance on sections (i.e. physical examination/ status) that contain more prevalent negated concepts.



Discussion re: Impact of section types

Explore 1. Train the same model on sections that contain more prevalent negated concepts
Explore 2. Include section type of each token as an input column .

- How does the base model perform on different sections?
- Why models perform worse on the highly-negated sections?
  - Smaller training set? —> downsample of base model for comparison