PyTAIL: Interactive and Incremental Learning of NLP Models with Human in the Loop for Online Data

Shubhanshu Mishra* (shubhanshu.com) - @TheShubhanshu
Jana Diesner (University of Illinois at Urbana-Champaign) - @janadiesner
*The work presented here was done during my PhD at UIUC

ArXiv: https://arxiv.org/abs/2211.13786
Dataset: https://doi.org/10.5281/zenodo.7236430
Code: https://github.com/socialmediaie/pytail
Video: https://www.youtube.com/watch?v=AwDu64gN8t4

Problem Formulation

Given a large unlabeled corpus, can we:

- label it efficiently using fewer human annotations?
- allow efficient human-in-the-loop injection of rules during the annotation process?
- update models efficiently to work with new data?

This setting is needed for social media data, where:

- Data is available in streaming mode, and
- Model should adapt to new data

Proposal

Given a large unlabeled corpus, can we:

- label it efficiently using fewer human annotations? ✅Active Learning
- allow efficient human-in-the-loop injection of rules during the annotation process? ✅Data and Rule suggestion interface
- update models efficiently to work with new data? ✅Online Learning

Scope: Classification Tasks for Social Media

Input

I know this tweet is late but I just want to say I absolutely fucking hated this season of
@GameOfThrones
what a waste of time.

Predict

Output

abusive

founta

abusive	hateful	normal	spam
0.830	0.084	0.085	0.002

waseem

none	racism	sexism
0.970	0.002	0.027

sentiment

	negative	neutral	positive
clarin	0.956	0.036	0.008
other	0.906	0.063	0.031
politics	0.917	0.048	0.035
semeval	0.966	0.030	0.004

uncertainity

sarcasm

not sarcasm	sarcasm
0.914	0.086

veridicality

definitely no	definitely yes	probably no	probably yes	uncertain
0.033	0.244	0.112	0.189	0.422

PyTAIL Benchmark of Active Learning on Social Media Text Classification

- Tasks for Social Media Text Classification: Abusive, Sentiment, Uncertainty
- 10 tasks, 200K social media posts
- To be released at: https://doi.org/10.5281/zenodo.7236430
- Derived from Social Media IE Multi Task Benchmark – https://doi.org/10.5281/zenodo.5867160

Data Stats

data	split	tokens	tweets	vocab
Airline	dev	20079	981	3273
	test	50777	2452	5630
	train	182040	8825	11697
Clarin	dev	80672	4934	15387
	test	205126	12334	31373
	train	732743	44399	84279
GOP	dev	16339	803	3610
	test	41226	2006	6541
	train	148358	7221	14342
Healthcare	dev	15797	724	3304
	test	16022	717	3471
	train	14923	690	3511
Obama	dev	3472	209	1118
	test	8816	522	2043
	train	31074	1877	4349
SemEval	dev	105108	4583	14468
	test	528234	23103	43812
	train	281468	12245	29673

Sentiment classification

data	split	tokens	tweets	vocab
Founta	dev	102534	4663	22529
	test	256569	11657	44540
	train	922028	41961	118349
WaseemSRW	dev	25588	1464	5907
	test	64893	3659	10646
	train	234550	13172	23042

Abusive content identification

data	split	tokens	tweets	vocab
Riloff	dev	2126	145	1002
	test	5576	362	1986
	train	19652	1301	5090
Swamy	dev	1597	73	738
	test	3909	183	1259
	train	14026	655	2921

Uncertainty indicator classification

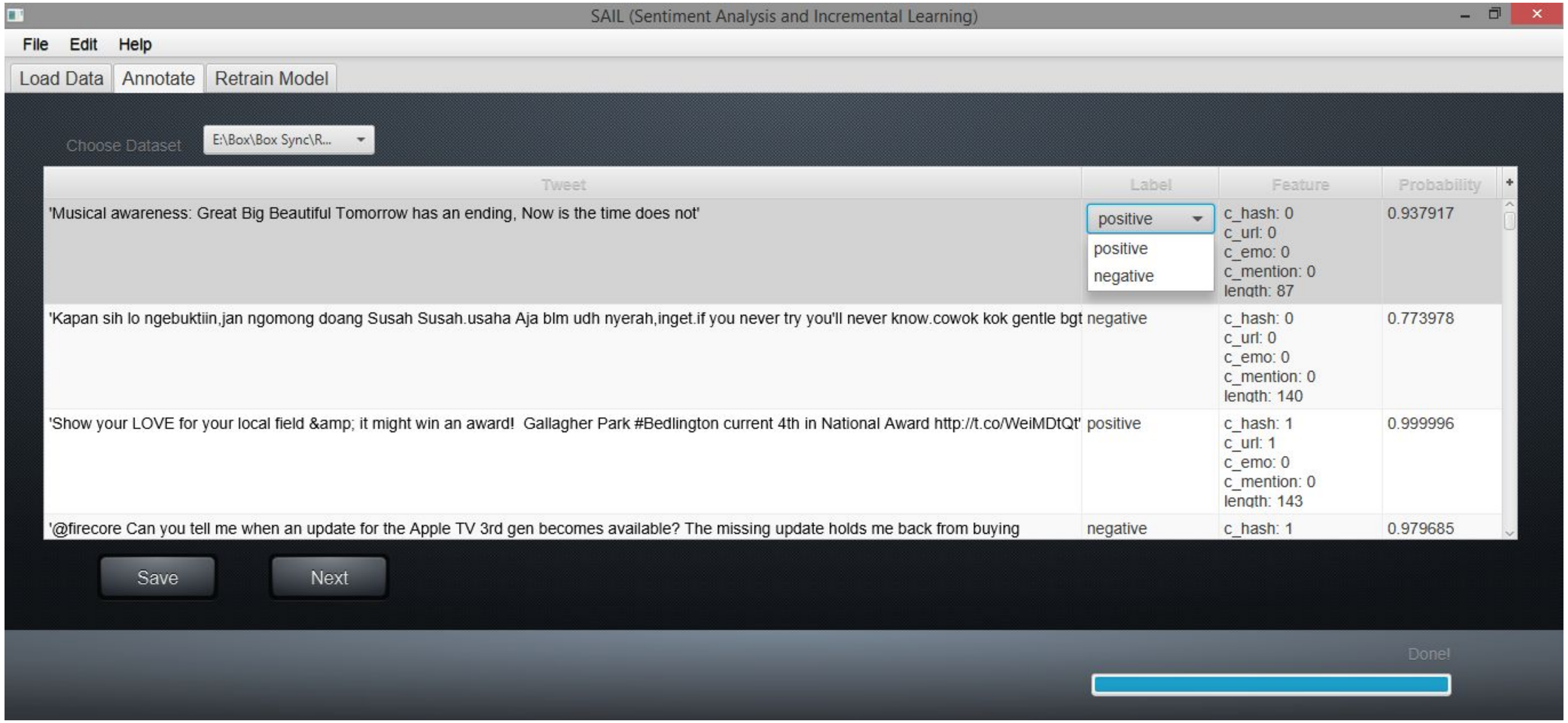
https://doi.org/10.5281/zenodo.5867160 and https://shubhanshu.com/phd_thesis/

SAIL: Sentiment Analysis with Incremental Human-in-the-Loop Learning and Lexical Resource Customization

- SAIL was written in Java and serves as a precursor for PyTAIL
- SAIL was written specifically for Sentiment Classification tasks and supports active online learning via SGD based updates.

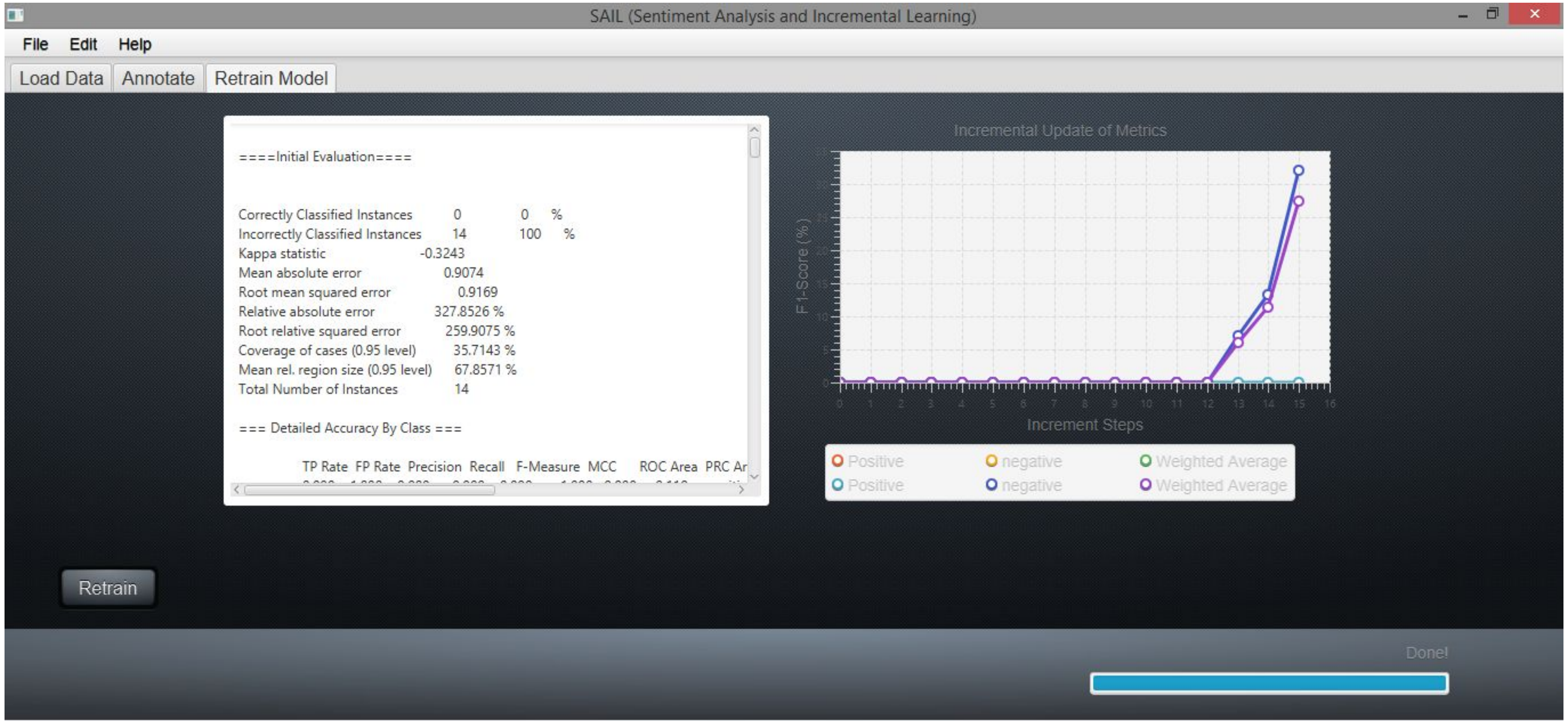
Mishra, Shubhanshu, Jana Diesner, Jason Byrne, and Elizabeth Surbeck. 2015. "Sentiment Analysis with Incremental Human-in-the-Loop Learning and Lexical Resource Customization." In *Proceedings of the 26th ACM Conference on Hypertext & Social Media - HT '15*, 323–25. New York, New York, USA: ACM Press. https://doi.org/10.1145/2700171.2791022.

SAIL: Sentiment Analysis with Incremental Human-in-the-Loop Learning and Lexical Resource Customization



Mishra, Shubhanshu, Jana Diesner, Jason Byrne, and Elizabeth Surbeck. 2015. "Sentiment Analysis with Incremental Human-in-the-Loop Learning and Lexical Resource Customization." In *Proceedings of the 26th ACM Conference on Hypertext & Social Media - HT '15*, 323–25. New York, New York, USA: ACM Press. https://doi.org/10.1145/2700171.2791022.

SAIL: Sentiment Analysis with Incremental Human-in-the-Loop Learning and Lexical Resource Customization

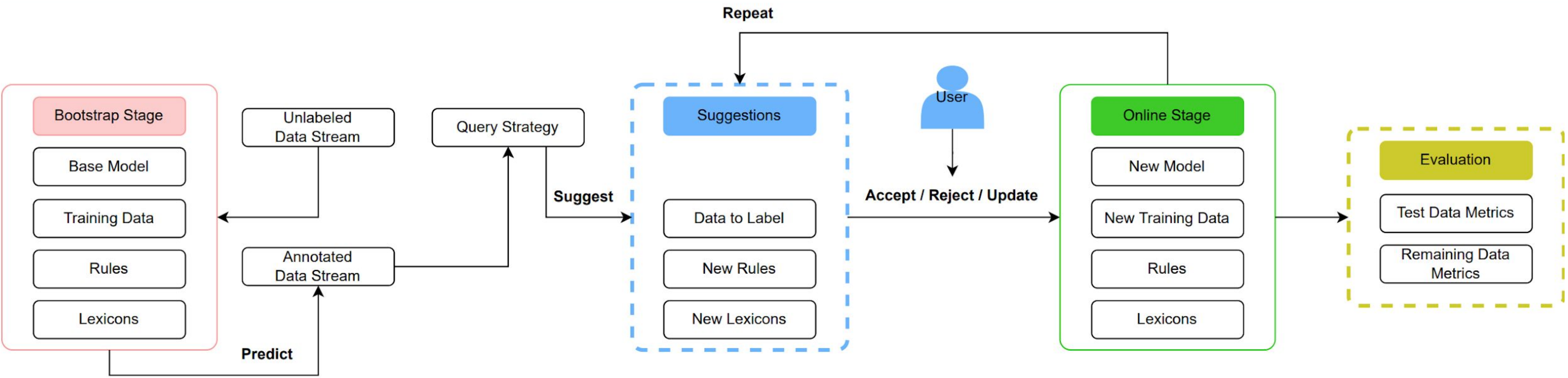


Mishra, Shubhanshu, Jana Diesner, Jason Byrne, and Elizabeth Surbeck. 2015. “Sentiment Analysis with Incremental Human-in-the-Loop Learning and Lexical Resource Customization.” In *Proceedings of the 26th ACM Conference on Hypertext & Social Media - HT ’15*, 323–25. New York, New York, USA: ACM Press. https://doi.org/10.1145/2700171.2791022.

PyTAIL Workflow

- Build an easy to use interface which allows users to perform human-in-the-loop annotation of data and incremental training of the model
- Enable injection of custom lexicons and rules for NLP application, with ability to suggest rules
- Support simulation mode to assess performance of active learning techniques
- Support human in the loop interface for interactive annotation and rule building
- Track performance of remaining data during simulation model to measure time to full annotation.
- Support different active learning algorithms
- Support different rule suggestion techniques

PyTAIL Workflow



Simulation and Human in the loop modes

Active Learning

1. Given a model and unlabeled data
2. Select samples from the unlabeled data to be annotated, based on selection criterion
3. Update model with collected labeled examples
4. Repeat steps 2 to 3 till desired accuracy is reached or data exhausted

Mishra et al. (2015)

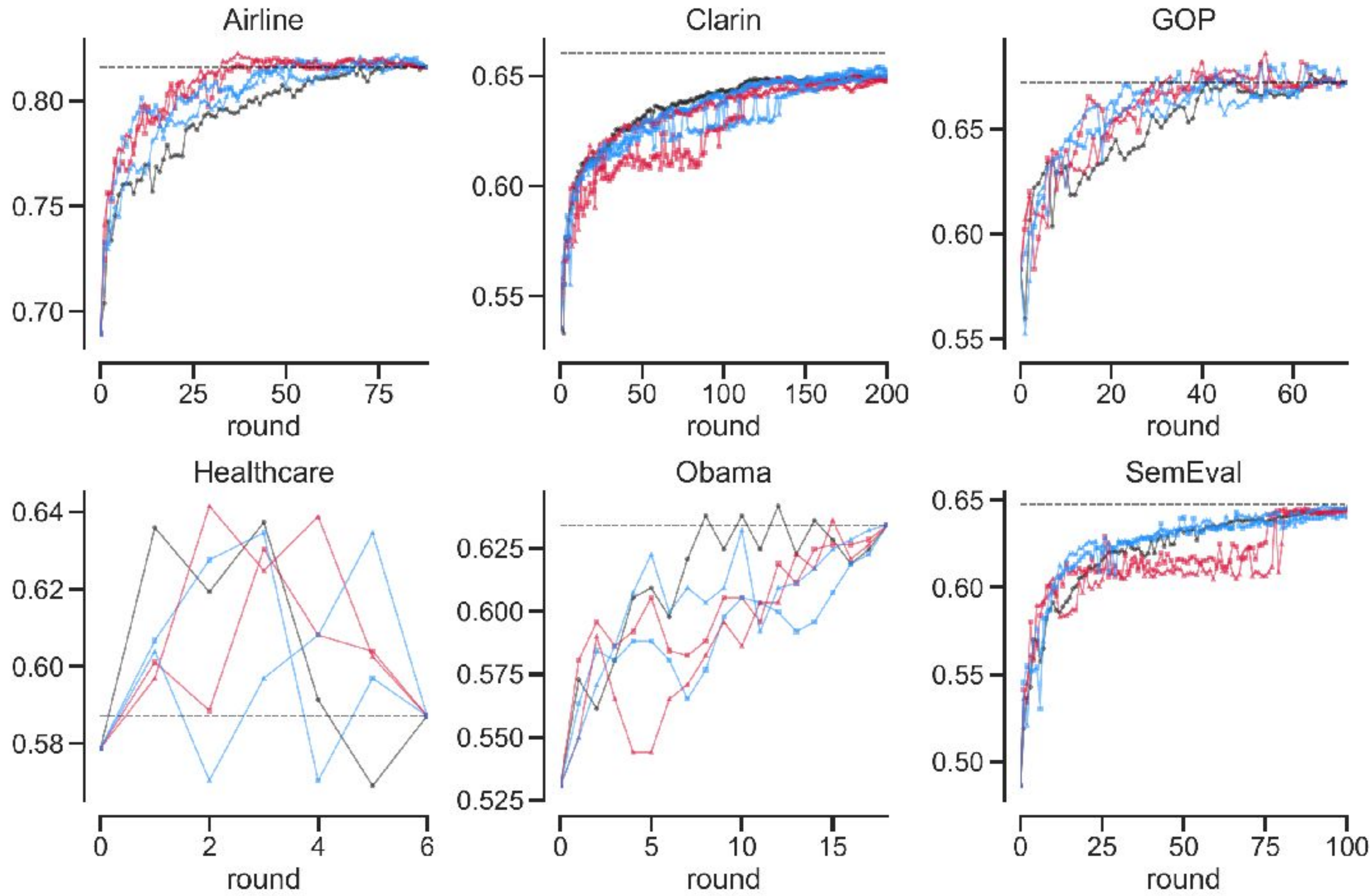
PyTAIL - API

```
class PyTAILTrainer:
    def __init__(
        model_fn,
        lexicon,
        rules,
        scoring_fn=entropy_scoring,
        selection_fn=select_top,
        simulation=True,
    ):
        # Define Trainer
        pass

    def update(suggestions):
        pass

    def select_new_data(model, unlabled_data):
        scores = scoring_fn(model, unlabled_data)
        new_training_data, new_rules, new_lexicons = selection_fn(unlabled_data, scores)
        if not simulation:
            new_training_data, new_rules, new_lexicons = ask_human(
                new_training_data, new_rules, new_lexicons
            )
        return new_training_data, new_rules, new_lexicons
```

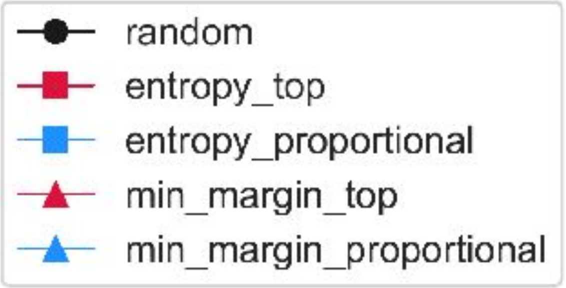
```
    def train_single_round(data):
        # Used for each active learning loop
        # Update model, lexicon, rules
        # request human judgement if simulation=False
        train_data = data[data.train]
        unlabled_data = data[~data.train]
        model = self.model_fn(train_data, self.lexicon, self.rules)
        if simulation:
            metrics = model.eval(unlabled_data)
        return model, metrics

    def train_multiple_rounds(
        data, seed_indices, per_round_budget, stopping_criteria: Callback
    ):
        # Run multiple rounds of PyTAIL learning
        data[seed_indices].train = True
        data[~seed_indices].train = False
        while stopping_criteria():
            model, metrics = self.train_single_round(data)
            suggestions = select_new_data(model, unlabled_data)
        self.update(suggestions)
        return all_metrics, base_metrics, training_indexes
```

Evaluation Workflow

- We evaluated PyTAIL simulation workflow on the PyTAIL benchmark
- Using a logistic regression model, and a continuously updated lexicon from the data
- The goal was the evaluate the performance of different active learning strategies
- We considered, random, entropy based, and min margin for candidate scoring.
- We considered top K and K sampled for candidate selection



- Each round query 100 samples
- Classifier is logistic regression with unigram and lexicon features
- Max rounds is 100 (except Clarin)

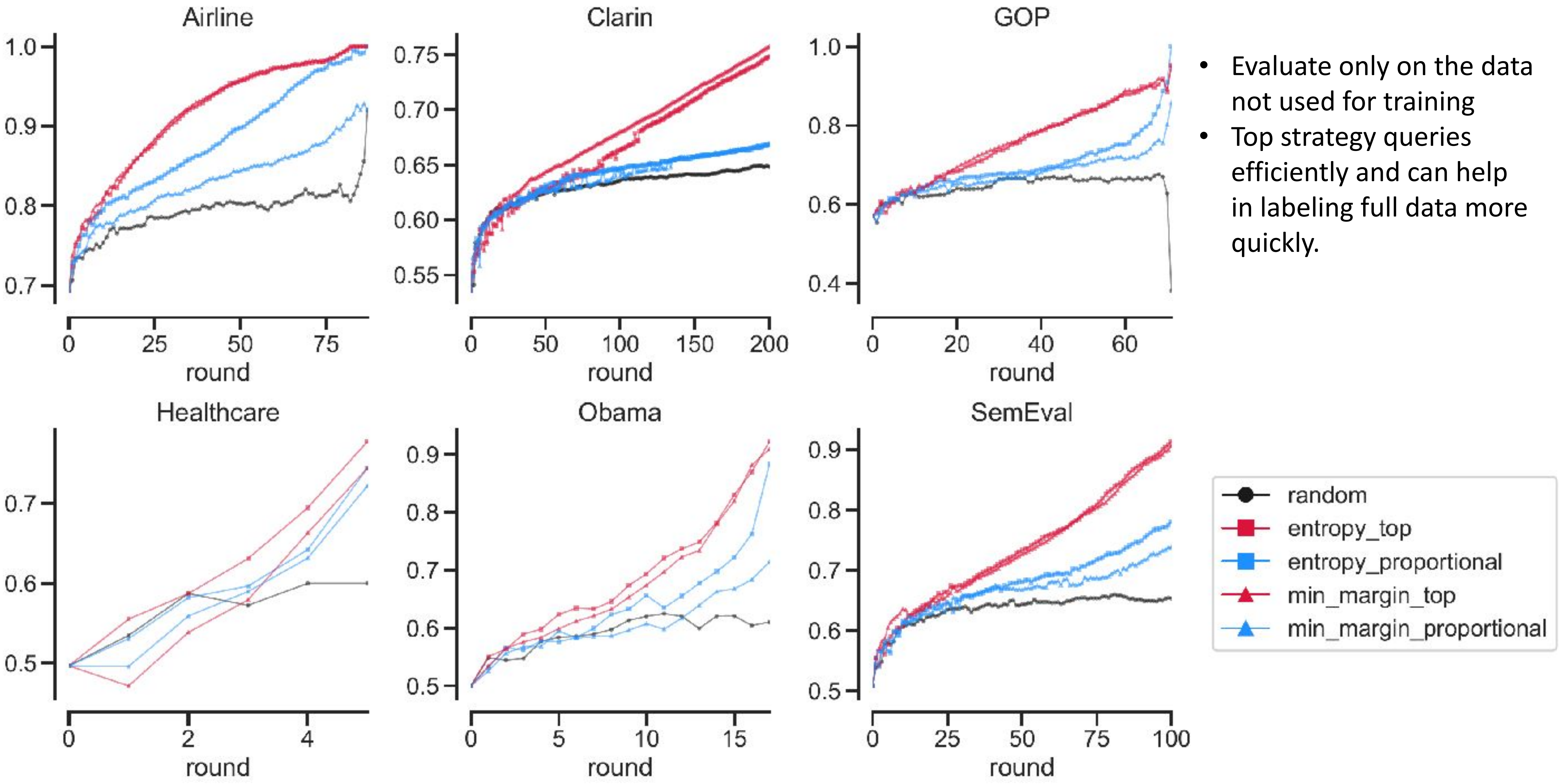


Data ordered alphabetically and X and Y axes are not shared.

https://github.com/socialmediaie/SocialMediaIE

Evaluation on remaining data

- Active learning systems only track performance on held out test set
- However, often goal is to quickly annotate a large unlabeled data
- We should hence track which methods quickly allows us to reach this goal by measuring the performance on the remaining data



- Evaluate only on the data not used for training
- Top strategy queries efficiently and can help in labeling full data more quickly.

Data ordered alphabetically and X and Y axes are not shared.

https://github.com/socialmediaie/SocialMediaIE

Benchmark Evaluation

Table 2: Performance of query strategies across datasets using around 10% training dataset.

task	dataset	round	N	N_{left}	$\%_{used}$	Full	Rand	E_{top}	E_{prop}	M_{top}	M_{prop}
				Test Dataset							
ABUSIVE	Founta	42	41,861	37,661	0.10	0.79	0.77	0.78	0.78	0.79	0.77
	WaseemSRW	14	13,072	11,672	0.11	0.82	0.79	0.78	0.77	0.78	0.76
SENTIMENT	Airline	9	8,725	7,825	0.10	0.82	0.76	0.78	0.79	0.77	0.77
	Clarin	45	44,299	39,799	0.10	0.66	0.63	0.61	0.62	0.63	0.63
	GOP	8	7,121	6,321	0.11	0.67	0.63	0.64	0.63	0.62	0.64
	Healthcare	1	590	490	0.17	0.59	0.64	0.60	0.61	0.60	0.60
	Obama	2	1,777	1,577	0.11	0.63	0.56	0.60	0.58	0.59	0.57
	SemEval	13	12,145	10,845	0.11	0.65	0.59	0.60	0.61	0.58	0.61
UNCERTAINITY	Riloff	2	1,201	1,001	0.17	0.78	0.77	0.76	0.77	0.76	0.79
	Swamy	1	555	455	0.18	0.39	0.39	0.40	0.39	0.34	0.31
				Remaining Dataset							
ABUSIVE	Founta	42	41,861	37,661	0.10	NaN	0.77	0.80	0.78	0.81	0.78
	WaseemSRW	14	13,072	11,672	0.11	NaN	0.78	0.79	0.77	0.80	0.76
SENTIMENT	Airline	9	8,725	7,825	0.10	NaN	0.75	0.79	0.79	0.80	0.78
	Clarin	45	44,299	39,799	0.10	NaN	0.62	0.62	0.62	0.64	0.63
	GOP	8	7,121	6,321	0.11	NaN	0.62	0.64	0.62	0.63	0.63
	Healthcare	1	590	490	0.17	NaN	0.53	0.56	0.53	0.47	0.50
	Obama	2	1,777	1,577	0.11	NaN	0.54	0.56	0.57	0.56	0.56
	SemEval	13	12,145	10,845	0.11	NaN	0.61	0.62	0.62	0.63	0.62
UNCERTAINITY	Riloff	2	1,201	1,001	0.17	NaN	0.80	0.82	0.84	0.82	0.81
	Swamy	1	555	455	0.18	NaN	0.37	0.40	0.40	0.33	0.36

- Our results show that Top K strategies lead to the fastest annotation of a given unlabeled corpora
- Random leads to the slowest annotation of the corpora.
- In terms of generalization capabilities most approaches are similar

Thank you

- Questions?
- Tweet to us at:
  - Shubhanshu Mishra - @TheShubhanshu
  - Jana Diesner - @janadiesner @DiesnerLab
- PyTAIL will be released soon at: https://github.com/socialmediaie/pytail
- Previous version of PyTAIL used for our experiments can be found as part of the SocialMediaIE tool: https://github.com/socialmediaie/SocialMediaIE/tree/master/SocialMediaIE/active_learning
- If you have questions or feature requests open an issue on GitHub at: https://github.com/socialmediaie/pytail/issues

References

- Mishra, Shubhanshu, Diesner, Jana, Byrne, Jason, & Surbeck, Elizabeth (2015). Sentiment Analysis with Incremental Human-in-the-Loop Learning and Lexical Resource Customization. In Proceedings of the 26th ACM Conference on Hypertext & Social Media - HT ’15 (pp. 323–325). New York, New York, USA: ACM Press. https://doi.org/10.1145/2700171.2791022
- Shubhanshu Mishra and Jana Diesner. 2022. PyTAIL: Interactive and Incremental Learning of NLP Models with Human in the Loop for Online Data. arXiv:2211.13786 [cs].