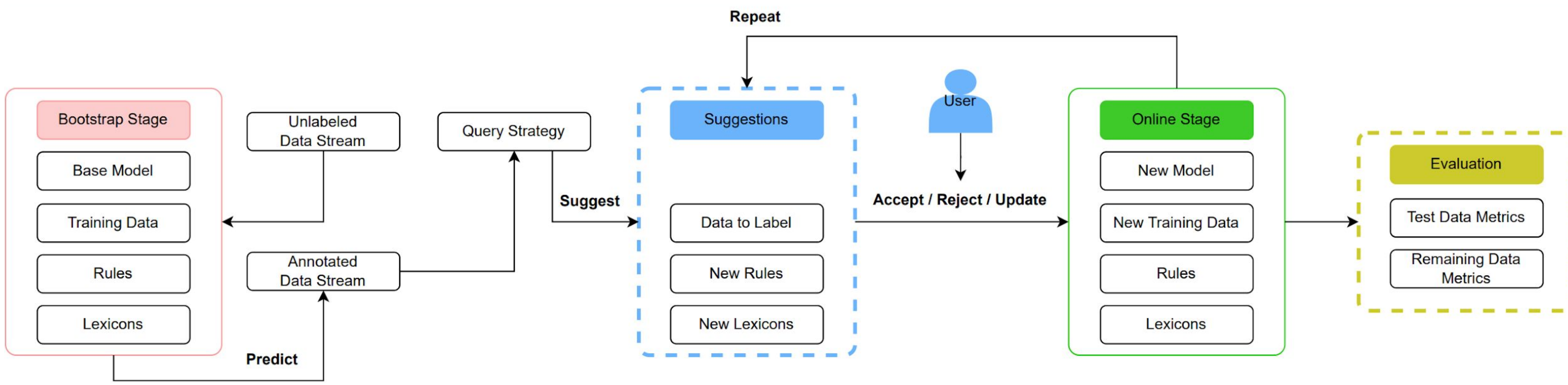
PyTAIL: Interactive and Incremental Learning of NLP Models with Human in the Loop for Online Data

NEURAL INFORMATION PROCESSING SYSTEMS

Shubhanshu Mishra* (shubhanshu.com), Jana Diesner (University of Illinois at Urbana-Champaign), *Work done while at UIUC.

https://github.com/socialmediaie/pytail | https://arxiv.org/abs/2211.13786



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?
- Proposal:
  - Use active learning for data labeling
  - Use interface to surface and inject prominent rules for efficient annotation
  - Use incremental learning algorithms for model updates
- Highly applicable to social media data:
  - Streaming data
  - Model should adapt to new data

PyTAIL Benchmark for Social Media Active Learning

- 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	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

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

Evaluation Workflow

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

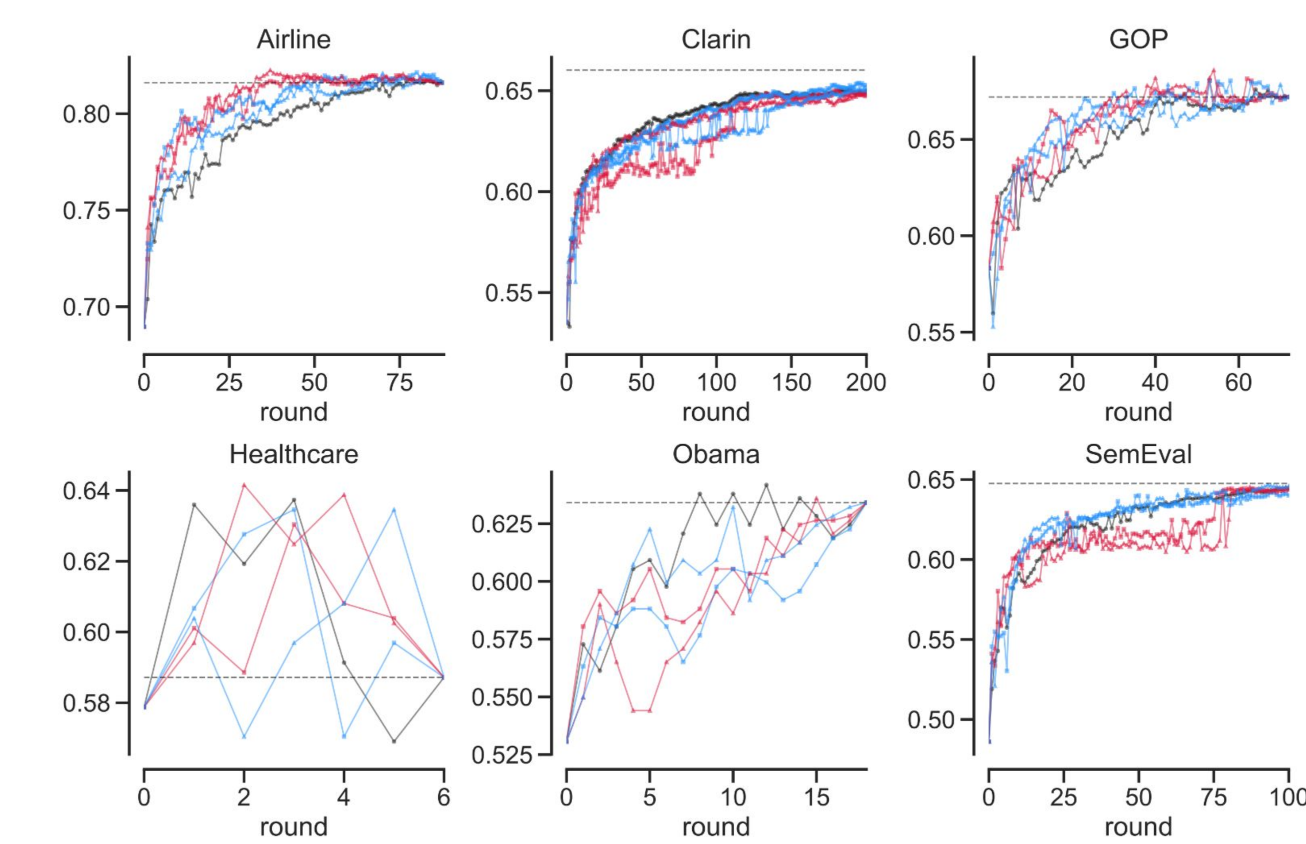
Resources

- ArXiv: https://arxiv.org/abs/2211.13786
- Dataset: https://doi.org/10.5281/zenodo.7236430
- Code: https://github.com/socialmediaie/pytail
- Video: https://youtu.be/AwDu64gN8t4

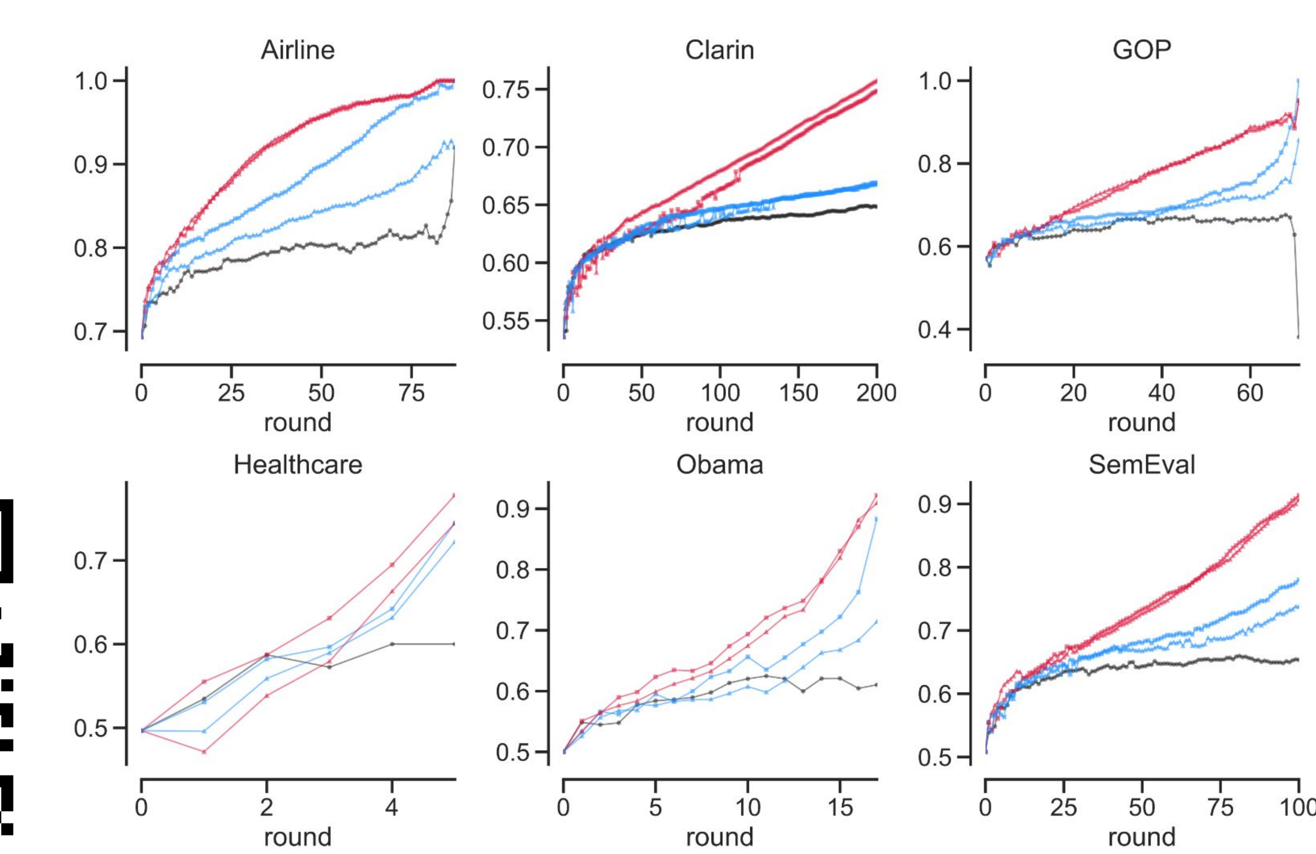
Evaluation of PyTAIL on benchmark

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
UNCERTAININITY	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
UNCERTAININITY	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



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



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