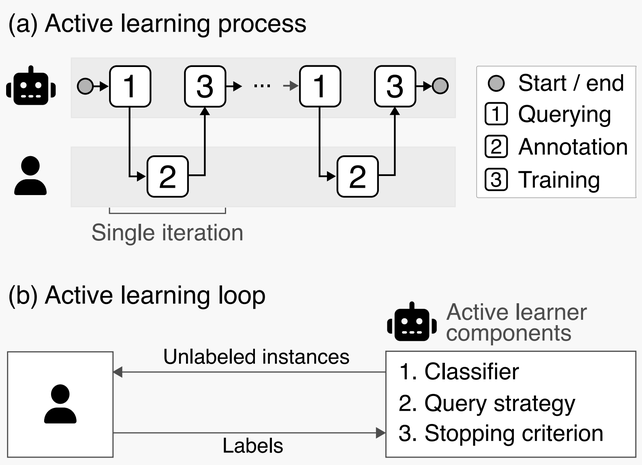
Small-Text: Active Learning for Text Classification in Python

Christopher Schröder, Lydia Müller, Andreas Niekler, Martin Potthast

Active Learning



Motivation

- Active learning experiments often involve a variety of strategies and therefore quickly become very complex.
- Existing active learning libraries rarely consider text classification and GPU-capable algorithms.

Contributions

- We provide an open source library for active learning for text classification.
- The library integrates `scikit-learn`, `PyTorch`, and `transformers`.
- Tried and tested components can be easily mixed and matched.
- In the experiment, we compare vanilla fine-tuning against contrastive learning-based fine-tuning with SetFit [1].

Software Features

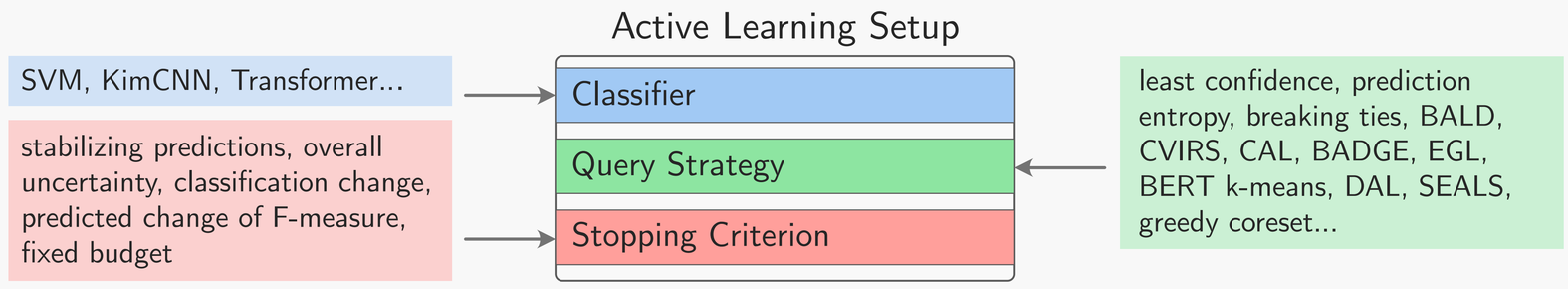
- State-of-the-art pool-based active learning for text classification.
- The library currently provides 14 query strategies and 5 stopping criteria.
- A modular architecture allows for a slim core installation (CPU) or an extended installaton (GPU).
- The extended installation offers one integration for the `PyTorch` and one for the `transformers` library.

Active Learning for Text Classification with unified interfaces for scikit-learn, PyTorch and transformers.

Github

github.com/webis-de/small-text

Quickly Build Experiments and Applications



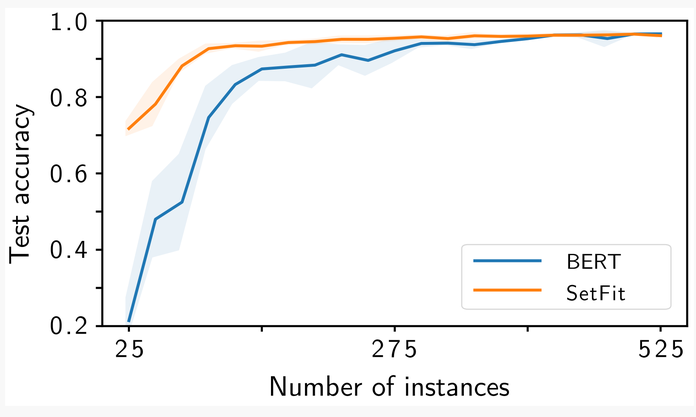
References and detailed information about each method can be found in the paper.

Comparison to Previous Software

Name	Active Learning			Code					
	QS	SC	Text Focus	GPU support	Unit Tests	Language	License	Last Update	Repository
`JCLAL`	18	2	✗	✗	✗	Java	GPL	2017	
`libact`	19	-	✗	✗	✓	Python	BSD-2-Clause	2021	
`modAL`	12	-	✗	✓	✓	Python	MIT	2022	
`ALiPy`	22	4	✗	✗	✓	Python	BSD-3-Clause	2022	
`BaaL`	9	-	✗	✓	✓	Python	Apache 2.0	2023	
`lrtc`	7	-	✓	✓	✗	Python	Apache 2.0	2021	
`scikit-activeml`	29	-	✗	✓	✓	Python	BSD-3-Clause	2023	
`ALToolbox`	19	-	✓	✓	✓	Python	MIT	2023	
`small-text`	14	5	✓	✓	✓	Python	MIT	2023	

A Github link and detailed information for each software can be found in the paper. The `low-resource-text-classification-framework` was abbreviated by `lrtc`.

Selected Results



(Dataset: TREC / Query Strategy: breaking ties)

Conclusions

- We introduce `small-text`, a modular Python library, which offers state-of-the-art active learning for text classification.
- `Small-text` has already been adopted in recent works [2, 3, 4] have already adopted `small-text` (and they also published their experiment code).
- Contrastive learning-based active learning is highly effective.

UNIVERSITÄT LEIPZIG

WEBIS.DE

[1] L. Tunstall, N. Reimers, U. E. S. Jo, L. Bates, D. Korat, M. Wasserblat, and O. Pereg, "Efficient few-shot learning without prompts," *arXiv:2209.11055*, 2022.
[2] H. Kirk, B. Vidgen, and S. Hale, "Is More Data Better? Re-thinking the Importance of Efficiency in Abusive Language Detection with Transformers-Based Active Learning," in *Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022)*, 2022, pp. 52–61.
[3] J. Gonsior, C. Falkenberg, S. Magino, A. Reusch, M. Thiele, and W. Lehner, "To softmax, or not to softmax: That is the question when applying active learning for transformer models," *arXiv:2210.03005*, 2022.
[4] J. Romberg and T. Escher, "Automated topic categorisation of citizens' contributions: Reducing manual labelling efforts through active learning," in *Electronic Government*, 2022, pp. 369–385.

This research was partially funded by the Development Bank of Saxony (SAB) under project numbers 100335729 and 100400221. Computations were done (in part) using resources of the Leipzig University Computing Centre. Printed at the Leipzig University Computing Centre.