Revisiting Uncertainty-based Query Strategies for Active Learning with Transformers

Christopher Schröder, Andreas Niekler, Martin Potthast

Motivation

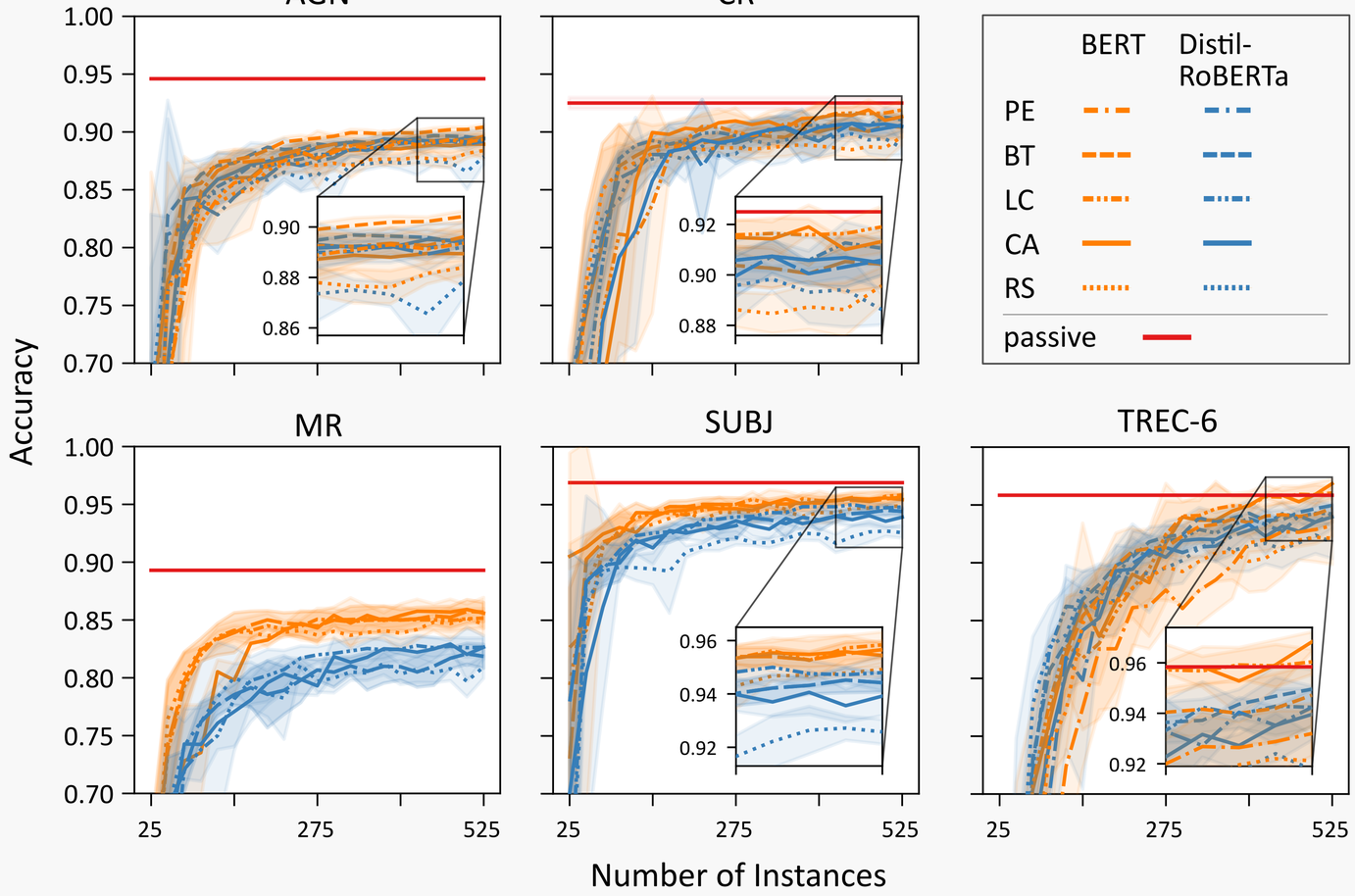
- Transformer models are increasingly used in active learning for text classification.
- Query strategies targeted at neural networks or text classification are computationally expensive.
- Uncertainty-based query strategies are computationally inexpensive but are usually considered only as a baseline.

Contributions

- We systematically investigate uncertainty-based query strategies in combination with transformer models (BERT [1], DistilRoBERTa [2]).
- Our experiments use five previously established but lately neglected text classification benchmarks.
- We investigate the effectiveness of using a transformer model with fewer parameters, DistiRoBERTa, for active learning.

Uncertainty-based query strategies with transformers are strong on text classification benchmarks.

Results: Learning Curves



Query Strategies

Prediction Entropy [3, 4]	$\underset{x_i}{\operatorname{argmax}} \left[-\sum_{j=1}^{c} P(y_i = j \mid x_i) \log P(y_i = j \mid x_i) \right]$
Breaking Ties [5, 6]	$\underset{x_i}{\operatorname{argmin}} \left[P(y_i = k_1^* \mid x_i) - P(y_i = k_2^* \mid x_i) \right]$
Least Confidence [7]	$\underset{x_i}{\operatorname{argmax}} \left[1 - P(y_i = k_1^* \mid x_i) \right]$
Contrastive Active Learning [8]	$\underset{x_i}{\operatorname{argmax}} \left[\frac{1}{m} \sum_{j=1}^{m} \mathrm{KL}(P(y_j \mid x_j^{knn}) \parallel P(y_i \mid x_i)) \right]$
Random	Sample i.i.d. from the unlabeled pool.

Data

Dataset Name (ID)	Type	Classes	Training	Test
AG's News (AGN)	N	4	120,000	7,600
Customer Reviews (CR)	S	2	3,397	378
Movie Reviews (MR)	S	2	9,596	1,066
Subjectivity (SUBJ)	S	2	9,000	1,000
TREC-6 (TREC-6)	Q	6	5,500	500

Types: N=News, S=Sentiment, Q=Questions.

Results: Summary

Model	Strategy	Mean Rank		Mean Result	
		Acc.	AUC	Acc.	AUC
SVM	PE	1.80	2.60	0.764	0.663
	BT	**1.60**	**1.60**	**0.767**	**0.697**
	LC	3.00	2.60	0.751	0.672
	CA	5.00	5.00	0.667	0.593
	RS	3.00	2.60	0.757	0.686
KimCNN	PE	1.60	2.40	**0.818**	0.742
	BT	**1.60**	**2.00**	**0.818**	**0.750**
	LC	3.80	2.80	0.810	0.732
	CA	3.80	4.80	0.793	0.711
	RS	3.60	2.40	0.804	0.749
D.RoBERTa	PE	2.60	3.00	0.901	0.856
	BT	2.20	**1.80**	0.902	**0.864**
	LC	**1.40**	2.00	**0.904**	0.860
	CA	3.00	3.40	0.901	0.852
	RS	5.00	4.20	0.884	0.853
BERT	PE	2.40	2.40	0.909	0.859
	BT	**2.00**	**1.60**	0.914	**0.873**
	LC	2.20	3.80	**0.917**	0.866
	CA	2.80	2.60	0.916	0.872
	RS	5.00	4.00	0.899	0.861

Selected Results

- Using transformer models we reach considerably higher AUC scores compared to Zhang et al. [9].
- Active learning reaches scores very close (and even surpasses) previous state-of-the-art results, and our own passive classification.
- DistilRoBERTa reaches scores only slightly worse than BERT using about 25% of the parameters.

Conclusions

- We find that, contrary to common practice, prediction entropy seems not to always be the strongest baseline.
- DistilRoBERTa achieves results close to BERT while using only about 25% of the parameters.
- Breaking ties, which is equal in the binary setting, consistently outperforms prediction entropy in multi-class scenarios.

github.com/webis-de/ACL-22

UNIVERSITÄT LEIPZIG

WEBIS.DE

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding", in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019, pp. 4171–4186.

[2] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter", *arXiv preprint arXiv:1910.01108*, 2019.

[3] N. Roy and A. McCallum, "Toward optimal active learning through sampling estimation of error reduction", in *Proceedings of the Eighteenth International Conference on Machine Learning (ICML)*, 2001, pp. 441–448.

[4] G. Schohn and D. Cohn, "Less is more: Active learning with support vector machines", in *Proceedings of the Seventeenth International Conference on Machine Learning (ICML)*, 2000, pp. 839–846.

[5] T. Scheffer, C. Decomain, and S. Wrobel, "Active hidden markov models for information extraction", in *Proceedings of the 4th International Conference on Advances in Intelligent Data Analysis (IDA)*, 2001, pp. 309–318.

[6] T. Luo, K. Kramer, D. B. Goldgof, L. O. Hall, S. Samson, A. Remsen, and T. Hopkins, "Active learning to recognize multiple types of plankton", *Journal of Machine Learning Research (JMLR)*, pp. 589–613, 2005.

[7] A. Culotta and A. McCallum, "Reducing labeling effort for structured prediction tasks", in *Proceedings of the 20th National Conference on Artificial Intelligence*, 2005, pp. 746–751.

[8] K. Margatina, G. Vernikos, L. Barrault, and N. Aletras, "Active learning by acquiring contrastive examples", in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021, pp. 650–663.

[9] Y. Zhang, M. Lease, and B. C. Wallace, "Active discriminative text representation learning", in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017, pp. 3386–3392.

This research was partially funded by the Development Bank of Saxony (SAB) under project number 100335729. Printed at the Leipzig University Computing Centre.