Revisiting Uncertainty-based Query Strategies for Active Learning with Transformers

Findings of ACL 2022

Paper and Code
github.com/webis-de/ACL-22

Christopher Schröder
Andreas Niekler
Martin Potthast
Introduction

**Active Learning**: minimize the labeling costs of training data acquisition while maximizing a model's performance (increase) with each newly labeled problem instance.
This Paper

Motivation

- Research has started to investigate transformer models ("transformers") for active learning but previous findings may not generalize to transformer models.
- Query strategies targeted at neural networks or text classification are computationally expensive.
- Uncertainty-based query strategies are (computationally inexpensive but) usually considered only as a baseline.

Contributions

- Systematic investigation of uncertainty-based query strategies paired with transformers.
- Evaluation on a five well-known lately neglected text classification benchmarks.
- We investigate the effectiveness of using a transformer model with fewer parameters, DistiRoBERTa, for active learning.
Experiment

**Models:** BERT [Devlin et al. 2019], DistilRoBERTA [Sanh et al. 2019] (and KimCNN [Kim 2014], SVM)

**Query Strategies:**

**Prediction Entropy**
[Roy and McCallum 2001; Schohn and Cohn 2000]

\[
\text{argmax}_{x_i} \left[ - \sum_{j=1}^{c} P(y_i = j| x_i) \log P(y_i = j| x_i) \right]
\]

**Breaking Ties**
[Scheffer et al. 2001; Luo et al. 2005]

\[
\text{argmin}_{x_i} \left[ P(y_i = k_1^* | x_i) - P(y_i = k_2^* | x_i) \right]
\]

**Least Confidence**
[Culotta and McCallum 2005]

\[
\text{argmax}_{x_i} \left[ 1 - P(y_i = k_1^* | x_i) \right]
\]

**Contrastive Active Learning**
[Margatina et al. 2021]

\[
\text{argmax}_{x_i} \left[ \frac{1}{m} \sum_{j=1}^{m} \text{KL}(P(y_j| x_j^{knn}) \parallel P(y_i| x_i)) \right]
\]

**Random Sampling**

Sample i.i.d. from the unlabeled pool.
## Experiment: Datasets

<table>
<thead>
<tr>
<th>Dataset Name (ID)</th>
<th>Type</th>
<th>Classes</th>
<th>Training</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>AG’s News (AGN) [Zhang et al. 2015]</td>
<td>News</td>
<td>4</td>
<td>120,000</td>
<td>(*) 7,600</td>
</tr>
<tr>
<td>Customer Reviews (CR) [Hu and Liu 2004]</td>
<td>Sentiment</td>
<td>2</td>
<td>3,397</td>
<td>378</td>
</tr>
<tr>
<td>Movie Reviews (MR) [Pang and Lee 2005]</td>
<td>Sentiment</td>
<td>2</td>
<td>9,596</td>
<td>1,066</td>
</tr>
<tr>
<td>Subjectivity (SUBJ) [Pang and Lee 2004]</td>
<td>Sentiment</td>
<td>2</td>
<td>9,000</td>
<td>1,000</td>
</tr>
<tr>
<td>TREC-6 (TREC-6) [Li and Roth 2002]</td>
<td>Questions</td>
<td>6</td>
<td>5,500</td>
<td>(*) 500</td>
</tr>
</tbody>
</table>

(*): Predefined test sets were available and adopted.
Evaluation: Learning Curves

Number of Instances

Accuracy

AGN
CR
MR
SUBJ
TREC-6

BERT
Distil-RoBERTa
PE
BT
LC
CA
RS
passive

0.70
0.75
0.80
0.85
0.90
0.95
1.00

25 275 525

©Christopher Schröder 2022
<table>
<thead>
<tr>
<th>Model</th>
<th>Strategy</th>
<th>Mean Rank</th>
<th>Mean Result</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Acc.</td>
<td>AUC</td>
</tr>
<tr>
<td>SVM</td>
<td>PE</td>
<td>1.80</td>
<td>2.60</td>
</tr>
<tr>
<td></td>
<td>BT</td>
<td><strong>1.60</strong></td>
<td><strong>1.60</strong></td>
</tr>
<tr>
<td></td>
<td>LC</td>
<td>3.00</td>
<td>2.60</td>
</tr>
<tr>
<td></td>
<td>CA</td>
<td>5.00</td>
<td>5.00</td>
</tr>
<tr>
<td></td>
<td>RS</td>
<td>3.00</td>
<td>2.60</td>
</tr>
<tr>
<td>KimCNN</td>
<td>PE</td>
<td>1.60</td>
<td>2.40</td>
</tr>
<tr>
<td></td>
<td>BT</td>
<td><strong>1.60</strong></td>
<td><strong>2.00</strong></td>
</tr>
<tr>
<td></td>
<td>LC</td>
<td>3.80</td>
<td>2.80</td>
</tr>
<tr>
<td></td>
<td>CA</td>
<td>3.80</td>
<td>4.80</td>
</tr>
<tr>
<td></td>
<td>RS</td>
<td>3.60</td>
<td>2.40</td>
</tr>
<tr>
<td>D.RoBERTa</td>
<td>PE</td>
<td>2.60</td>
<td>3.00</td>
</tr>
<tr>
<td></td>
<td>BT</td>
<td>2.20</td>
<td><strong>1.80</strong></td>
</tr>
<tr>
<td></td>
<td>LC</td>
<td><strong>1.40</strong></td>
<td><strong>2.00</strong></td>
</tr>
<tr>
<td></td>
<td>CA</td>
<td>3.00</td>
<td>3.40</td>
</tr>
<tr>
<td></td>
<td>RS</td>
<td>5.00</td>
<td>4.20</td>
</tr>
<tr>
<td>BERT</td>
<td>PE</td>
<td>2.40</td>
<td>2.40</td>
</tr>
<tr>
<td></td>
<td>BT</td>
<td><strong>2.00</strong></td>
<td><strong>1.60</strong></td>
</tr>
<tr>
<td></td>
<td>LC</td>
<td>2.20</td>
<td>3.80</td>
</tr>
<tr>
<td></td>
<td>CA</td>
<td>2.80</td>
<td>2.60</td>
</tr>
<tr>
<td></td>
<td>RS</td>
<td>5.00</td>
<td>4.00</td>
</tr>
</tbody>
</table>

- Surprisingly: prediction entropy is outperformed by breaking ties.
- For DistilRoBERTa: least confidence also outperforms prediction entropy.
- DistilRoBERTa performs only slightly worse than BERT
Evaluation: Further Results

- Using transformer models we reach considerably higher AUC scores compared to Zhang et al. (2017).

- Active learning is very close (and even surpasses) previous state-of-the-art results, and our own passive classification, in terms of final accuracy (using a fraction of the data).

- Detailed results and runtimes per configuration are reported in the paper’s appendix.
Conclusion

Experiment: Active Learning for Text Classification
- BERT, DistilRoBERTa
- Several sentence classification datasets
- Four query strategies and a baseline

Findings
- The supposedly strongest baseline, prediction entropy, “is not so strong”.
- Breaking ties consistently outperforms prediction entropy in multi-class scenarios.
- DistilRoBERTa achieves results close to BERT while using only about 25% of the parameters.
Conclusion

Experiment: Active Learning for Text Classification

- BERT, DistilRoBERTa
- Several sentence classification datasets
- Four query strategies and a baseline

Findings

- The supposedly strongest baseline, prediction entropy, “is not so strong”.
- Breaking ties consistently outperforms prediction entropy in multi-class scenarios.
- DistilRoBERTa achieves results close to BERT while using only about 25% of the parameters.

Thank you!


In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pages 168–177.
Yoon Kim. 2014. 
Convolutional neural networks for sentence classification. 

Active learning to recognize multiple types of plankton. 

Xin Li and Dan Roth. 2002. 
Learning question classifiers. 

Active learning by acquiring contrastive examples. 
A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts.  
In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL), pages 271–278.

Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales.  

Nicholas Roy and Andrew McCallum. 2001.  
Toward optimal active learning through sampling estimation of error reduction.  


Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015.
Character-level convolutional networks for text classification.