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

**Active Learning**

(a) Active learning process

1. Start / end
2. Querying
3. Annotation
4. Training

(b) Active learning loop

1. Active learner components
   - Classifier
   - Query strategy
   - Stopping criterion

**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 installation (GPU).
- The extended installation offers one integration for the PyTorch and one for the transformers library.

**Quickly Build Experiments and Applications**

**Active Learning Setup**

- **Classifier**
  - SVM, KimCNN, Transformer...
  - least confidence, prediction entropy, breaking ties, BALD, k-means, DAL, SEALS, greedy coreset...

- **Query Strategy**
  - stabilizing predictions, overall uncertainty, classification change, fixed budget

- **Stopping Criterion**
  - predicted change of F-measure, fixed budget

**Comparison to Previous Software**

<table>
<thead>
<tr>
<th>Name</th>
<th>Active Learning</th>
<th>Code</th>
<th>Last Update</th>
<th>Repository</th>
</tr>
</thead>
<tbody>
<tr>
<td>JCLAL</td>
<td>QS 18 SC 2 Text</td>
<td>Pythons</td>
<td>2017</td>
<td>GitHub</td>
</tr>
<tr>
<td>libact</td>
<td>19 - X</td>
<td>Python</td>
<td>2021</td>
<td>GitHub</td>
</tr>
<tr>
<td>modAL</td>
<td>12 - X</td>
<td>Python</td>
<td>2022</td>
<td>GitHub</td>
</tr>
<tr>
<td>ALPy</td>
<td>22 - 4</td>
<td>Python</td>
<td>2022</td>
<td>GitHub</td>
</tr>
<tr>
<td>BaaL</td>
<td>9 - X</td>
<td>Python</td>
<td>2023</td>
<td>GitHub</td>
</tr>
<tr>
<td>lrtc</td>
<td>7 - X</td>
<td>Python</td>
<td>2021</td>
<td>GitHub</td>
</tr>
<tr>
<td>scikit-activeml</td>
<td>29 - X</td>
<td>Python</td>
<td>2023</td>
<td>GitHub</td>
</tr>
<tr>
<td>ALToolbox</td>
<td>19 - X</td>
<td>Python</td>
<td>2023</td>
<td>GitHub</td>
</tr>
<tr>
<td>small-text</td>
<td>14 5</td>
<td>Python</td>
<td>2023</td>
<td>GitHub</td>
</tr>
</tbody>
</table>

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**

<table>
<thead>
<tr>
<th>Test accuracy</th>
<th>Number of instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>25</td>
</tr>
<tr>
<td>0.4</td>
<td>50</td>
</tr>
<tr>
<td>0.6</td>
<td>75</td>
</tr>
<tr>
<td>0.8</td>
<td>100</td>
</tr>
<tr>
<td>1.0</td>
<td>125</td>
</tr>
</tbody>
</table>

(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] and has already adopted small-text (and they also published their experiment code).
- Contrastive learning-based active learning is highly effective.

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

---

**Notes**