Active Learning for Entity Filtering in Microblog Streams

The Entity Filtering Task
Filter out tweets that are not related to a given entity of interest

Active Learning
User feedback for updating the classification model

Setup
- Binary classification problem
- Support vector machines
- Simulated feedback from ground truth
- RepLab 2013 dataset
  http://nlp.uned.es/replab2013
- Evaluation metrics: Accuracy, Reliability & Sensitivity

Results
Passive Learning Baseline

<table>
<thead>
<tr>
<th>Run</th>
<th>Accuracy</th>
<th>F(R,S)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best system</td>
<td>0.91</td>
<td>0.49</td>
</tr>
<tr>
<td>RepLab 2013</td>
<td>0.92</td>
<td>0.46</td>
</tr>
</tbody>
</table>

Effectiveness of Active Learning
Margin Sampling significantly outperforms Random Sampling

Outliers
Margin Sampling performs significantly better than all density approaches

Initial Training Reduction
The cost of training the initial model can be substantially reduced

Conclusions
- Active learning scenario for entity filtering is feasible
- Less annotation is needed when annotation is done on the fly
- Compared state-of-the-art sampling methods: margin sampling works best
- Code available at http://damiano.github.io/al-ef