IMMM 2020
Towards Inter-Rater-Agreement-Learning
Michael Spranger
How AI shape our Life

It is difficult to think of a major industry that AI will not transform. This includes healthcare, education, transportation, retail, communications, and agriculture. There are surprisingly clear paths for AI to make a big difference in all of these industries.
Andrew Ng

It's very clear that AI is going to impact every industry. I think that every nation needs to make sure that AI is a part of their national strategy. Every country will be impacted.
Jensen Huang

I think that AI will lead to a low cost and better quality life for millions of people. Like electricity, it's a possibility to build a wonderful society.
Andrew Ng
A good AI’s needs

Commonly, a human-labeled dataset is considered as ground-truth.

The truth is rarely pure and never simple.
*Oscar Wilde*

Ideally, an expert-labeled dataset should be considered as ground-truth.

If you have a lot of data and you want to create value from that data, one of the things you might consider is building up an AI team.
*Andrew Ng*

Usually, machine learning needs much data, but there are not enough experts to label it.
Consensus by majority

I vote for COFFEE.

Let's have COFFEE.

What about TEA?

I prefer TEA.

COFFEE is better.

COFFEE = 3
TEA = 2

https://www.google.com/search?q=consensus&tbs=isch&hl=de&lr=lang&client=opera8&hs=LFj&sa=X&ved=8CAE&source=TCMIsp7uS&ses=CFQAAAAAdAAAAABAF&hl=de&gl=de&client=firefox-b&rlz=1C5CHFA_enDE100SDE100&ei=5ZQjX7atlAXJ8QZ1o5cBAQ
Are all ratings equally valuable?

Should all ratings for an item have the same weight?
# Weighted Learning Approach

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R_j$</td>
<td>rater’s competence</td>
</tr>
<tr>
<td>$f(t_{ji}, C_j)$</td>
<td>value depending on the response time and the conscientiousness of a specific rater who needs to annotate an item $i$ at time $t$</td>
</tr>
<tr>
<td>$\beta_i$</td>
<td>weighting parameter</td>
</tr>
<tr>
<td>$x_i$</td>
<td>feature (self-judgement, Intra-Rater-Agreement,…)</td>
</tr>
</tbody>
</table>

\[
W_{ji} = R_j - f(t_{ji}, C_j)
\]

\[
R_j = \frac{\sum_{l=1}^{n} \beta_l x_{lj}}{\frac{1}{|J|} \sum_{j'=1}^{|J|} \sum_{l=1}^{n} \beta_l x_{lj'}}
\]

\[
f(t_{ji}) = \begin{cases} 
0, & C_j > \bar{t}_i \land t_{ji} \in [\bar{t}_i, C_j] \\
0, & C_j < \bar{t}_i \land t_{ji} \in [C_j, \bar{t}_i] \\
t_{ji} - \bar{t}_i, & C_j > \bar{t}_i \land t_{ji} \notin [\bar{t}_i, C_j] \\
\bar{t}_i - t_{ji}, & C_j < \bar{t}_i \land t_{ji} \notin [C_j, \bar{t}_i]
\end{cases}
\]
Weighted Learning Approach

Taking time (already labeled items) into account:

\[
w_{ji} = (1 - b_{ji})R_j - f(t_{ji}, C_j) + \frac{1}{b_{ij}} \sum w_{j(i-1)}
\]

\[
b_{ji} = \frac{1}{\lambda(i-1)} \sum_{k=0}^{i-1} \begin{cases} 1, & \text{for } j \text{ in majority for item } k \\ 0, & \text{else} \end{cases}
\]
### Preliminary Results

**3000 texts from music domain**

<table>
<thead>
<tr>
<th>Threshold</th>
<th>is Music</th>
<th></th>
<th>Uncertain</th>
<th></th>
<th>not Music</th>
<th></th>
<th>No Majority</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>UW</td>
<td>W</td>
<td>UW</td>
<td>W</td>
<td>UW</td>
<td>W</td>
<td>UW</td>
<td>W</td>
</tr>
<tr>
<td>0.5</td>
<td>1534</td>
<td>1560</td>
<td>3</td>
<td>6</td>
<td>1344</td>
<td>1357</td>
<td>119</td>
<td>77</td>
</tr>
<tr>
<td>0.55</td>
<td>1456</td>
<td>1481</td>
<td>2</td>
<td>2</td>
<td>1298</td>
<td>1302</td>
<td>244</td>
<td>215</td>
</tr>
<tr>
<td>0.6</td>
<td>1373</td>
<td>1410</td>
<td>2</td>
<td>1</td>
<td>1240</td>
<td>1249</td>
<td>386</td>
<td>339</td>
</tr>
<tr>
<td>0.65</td>
<td>1278</td>
<td>1314</td>
<td>0</td>
<td>0</td>
<td>1175</td>
<td>1182</td>
<td>547</td>
<td>504</td>
</tr>
<tr>
<td>0.7</td>
<td>1128</td>
<td>1191</td>
<td>0</td>
<td>0</td>
<td>1115</td>
<td>1139</td>
<td>757</td>
<td>670</td>
</tr>
<tr>
<td>0.75</td>
<td>982</td>
<td>1072</td>
<td>0</td>
<td>0</td>
<td>1064</td>
<td>1081</td>
<td>954</td>
<td>847</td>
</tr>
<tr>
<td>0.8</td>
<td>825</td>
<td>918</td>
<td>0</td>
<td>0</td>
<td>998</td>
<td>1025</td>
<td>1177</td>
<td>1057</td>
</tr>
<tr>
<td>0.85</td>
<td>648</td>
<td>739</td>
<td>0</td>
<td>0</td>
<td>931</td>
<td>952</td>
<td>1421</td>
<td>1309</td>
</tr>
<tr>
<td>0.90</td>
<td>435</td>
<td>504</td>
<td>0</td>
<td>0</td>
<td>821</td>
<td>842</td>
<td>1744</td>
<td>1654</td>
</tr>
<tr>
<td>0.95</td>
<td>236</td>
<td>270</td>
<td>0</td>
<td>0</td>
<td>601</td>
<td>636</td>
<td>2163</td>
<td>2094</td>
</tr>
</tbody>
</table>

„No Majority“ decreases for each threshold
Conclusion & Future Work

• flexible weighting approach for Inter-Rater-Agreement
• strengths and weaknesses of different raters are considered
• automatic adaptation to dynamic user characteristics like concentration, motivation etc.
• results on a first dataset providing only few parameters leads to less items with “no majority”
• Future work will incorporating tests on a multi-lingual dataset including more features