The BERT probabilities of the tokens agree and disagree are correlated and we can exploit it.

How Gender Interacts with Political Values
Adnan Al Ali and Jindřich Libovický
adnan@matfyz.cz  libovicky@ufal.mff.cuni.cz
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Prague, Czech Republic

Motivation

• Pre-training corpora often contain unmoderated content.
• Gender bias and political values: how do they interact?

Methodology

• We work with the logarithms of the probabilities.
• Agree is usually rated higher than disagree, regardless of the statement.
• The (log-)probabilities for agree and disagree are correlated.
• We introduce apolitical calibration data to estimate the best fit line and variance of the error w.r.t. the line.
• Ratings below and above the line are considered as disagreeing and agreeing respectively.
• The exact rating is calculated using the CDF for $\mathcal{N}(0, \sigma^2)$.

Results

<table>
<thead>
<tr>
<th>Model</th>
<th>AntiAuth</th>
<th>CultLib</th>
<th>EconEq</th>
<th>Trib</th>
<th>Average Rating</th>
<th>Std Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>FERNET</td>
<td>3.8</td>
<td>3.9</td>
<td>3.8</td>
<td>3.8</td>
<td>2.9</td>
<td>0.5</td>
</tr>
<tr>
<td>XLM-R</td>
<td>3.2</td>
<td>3.3</td>
<td>3.2</td>
<td>3.2</td>
<td>3.3</td>
<td>0.6</td>
</tr>
<tr>
<td>mBERT</td>
<td>3.4</td>
<td>3.5</td>
<td>3.5</td>
<td>3.5</td>
<td>3.6</td>
<td>0.7</td>
</tr>
</tbody>
</table>

• We compare the obtained ratings to a real-life political values study of polled on Czech-speaking people, divided by gender.
• Most models made little distinction between the masculine and the feminine sentences, although the ratings differ in the real-life data
• All models underestimated the rating of cultural liberalism.
• All models overestimated the rating of economic equity.
• mBERT had the strongest opinions.
• Many ratings are close to the midpoint of the scale, with a large variance.

Conclusions

Most models made little to no distinction between the feminine and the masculine sentences.
Most models rated the sentences corresponding to the same value inconsistently, leading to a large variance.

We did not find any significant systematic perceived political values in the models.