**INTRODUCTION**

- NLP Shared Task for automatic detection of offensive language in Twitter
- OffensEval 2020 provides test set along with a weakly labeled large set of ca. 9 million Tweets for training
- OffensEval 2019 training data is still valid to use since it is the base for the weakly labeled training data
- We use 2020 training data for unsupervised pre-training and 2019 training + test data for supervised training

Two main contributions:

2. Study on further pre-training of the RoBERTa model with masked language for performance improvements. This model achieved the first place in the OffensEval 2020 Shared Task A for English.

**DATASET**

<table>
<thead>
<tr>
<th>Language</th>
<th>OFF</th>
<th>NOT</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>1,418</td>
<td>60,479</td>
<td>71,897</td>
</tr>
</tbody>
</table>

**EVALUATING TRANSFORMERS**

**Pretraining with masked language modeling (MLM)**

- BERT (Devlin et. al. 2019)
- RoBERTa (Liu et al. 2019)

**Multi-lingual masked language modeling**

- XLM-RoBERTa (Conneau et al. 2019)

**RESULTS**

<table>
<thead>
<tr>
<th>Model</th>
<th>NOT</th>
<th>F</th>
<th>P</th>
<th>R</th>
<th>F1</th>
<th>Macro F1</th>
<th>Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALBERT</td>
<td>0.966</td>
<td>0.987</td>
<td>0.988</td>
<td>0.988</td>
<td>0.988</td>
<td>0.988</td>
<td>0.988</td>
</tr>
<tr>
<td>BERT</td>
<td>0.964</td>
<td>0.986</td>
<td>0.986</td>
<td>0.986</td>
<td>0.986</td>
<td>0.986</td>
<td>0.986</td>
</tr>
<tr>
<td>RoBERTa</td>
<td>0.965</td>
<td>0.987</td>
<td>0.987</td>
<td>0.987</td>
<td>0.987</td>
<td>0.987</td>
<td>0.987</td>
</tr>
</tbody>
</table>

**EVALUATING TRANSFORMERS**

<table>
<thead>
<tr>
<th>Language</th>
<th>OFF</th>
<th>NOT</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>1,418</td>
<td>60,479</td>
<td>71,897</td>
</tr>
</tbody>
</table>

**CONCLUSIONS**

- Multiple pre-trained transformer architectures among the best performing models for the OffensEval 2020 Shared Task
- ALBERT ensemble performs best for Task A (with less model parameters than BERT or RoBERTa)
- Still room for improvement on Tasks B and C (e.g. through measures against negative effects from class imbalance)
- Weak labels for training are not helpful
- Further pre-training with MLM on unlabelled in-domain data improves offensive language detection

**ACKNOWLEDGEMENTS**

The work was supported by BWFG Hamburg (Germany) within the “Forum 4.0” project as part of the ahi digital funding line, and the DFG project “FAME” (WI 4949/2-1, grant no. 406289255).

**REFERENCES**


