Introduction

The poster presents various pre-training strategies that aid in improving the accuracy of the sentiment classification task for Latvian tweets. We experiment with existing language representation models along with in-domain data. The best results are achieved when pre-training the mBERT language representation model with in-domain data and introducing emoticons to the mBERT vocabulary during pre-training.

Datasets

The following datasets were used:

- **Gold** – a corpus consisting of 6777 human-annotated Latvian tweets from the period of August 2016 till November 2016.
- **Peisenieks** – a corpus consisting of 1178 human-annotated Latvian tweets created by Peisenieks and Skadinš.
- **Auto** – three sets of tweets from the period of August 2016 till July 2018 automatically annotated based on sentiment-identifying emoticons that are present in the tweets – 23,685 tweets with emoticons, 23,685 tweets with removed emoticons, and 47,370 tweets with both present and removed emoticons.
- **English** – a corpus of 45,530 various human-annotated English tweets from various sources that were machine-translated into Latvian.
- A time-balanced evaluation set that consists of 1000 tweets from the period of August 2016 till July 2018.
- Latvian tweets from the Latvian Tweet Corpus. The corpus consists of 4,640,804 unique Latvian tweets that have been collected during the timeframe from August 2016 till March 2020.

Methods

The following strategies were used:

**Pre-training**
- mBERT - vanilla version (Base).
- mBERT - pre-trained on the Latvian Tweet Corpus (Pre).
- mBERT - pre-trained on the Latvian Tweet Corpus plus emoticons are added to the vocabulary of mBERT (Pre+Emo).
- ALBERT and ELECTRA.

**Fine-tuning**

We use a 3 class-classification layer on top of the representations obtained from the model representation models listed above.

Error Analysis

Possible reasons of misclassification:

- 32% - world knowledge or external context needed for predicting the correct sentiment
- 17% - words of opposite sentiment
- 13% - sarcastic expressions
- 12% - multiple polarities in one tweet
- 4% - double negation
- 3% - spelling mistakes and lack of diacritic

Results

Table 1: Results of the classifier (Accuracy Scores).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Perceptron [1]</th>
<th>mBERT Base</th>
<th>mBERT Pre</th>
<th>mBERT Pre+Emo</th>
<th>ALBERT</th>
<th>ELECTRA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gold</td>
<td>0.661</td>
<td>0.678</td>
<td>0.756</td>
<td>0.754</td>
<td>0.661</td>
<td>0.711</td>
</tr>
<tr>
<td>Gold+Peisenieks</td>
<td>0.676</td>
<td>0.692</td>
<td>0.747</td>
<td>0.764</td>
<td>0.698</td>
<td>0.706</td>
</tr>
<tr>
<td>Gold+Auto (with ◦)</td>
<td>0.624</td>
<td>0.679</td>
<td>0.769</td>
<td>0.748</td>
<td>0.649</td>
<td>0.680</td>
</tr>
<tr>
<td>Gold+Auto (no ◦)</td>
<td>0.512</td>
<td>0.523</td>
<td>0.648</td>
<td>0.660</td>
<td>0.483</td>
<td>0.621</td>
</tr>
<tr>
<td>Gold+Auto (both)</td>
<td>0.487</td>
<td>0.526</td>
<td>0.618</td>
<td>0.657</td>
<td>0.509</td>
<td>0.364</td>
</tr>
<tr>
<td>Gold+English</td>
<td>0.613</td>
<td>0.698</td>
<td>0.692</td>
<td>0.720</td>
<td>0.669</td>
<td>0.684</td>
</tr>
</tbody>
</table>

Conclusion

Our experiments allowed us to achieve an accuracy increase by up to 13% over previous methods when pre-training word embedding models with in-domain unlabelled data and fine-tuning the models on relatively small supervised datasets.

Acknowledgements

The work presented in this paper has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 812997 and under the name CLEOPATRA (Cross-lingual Event-centric Open Analytics Research Academy). This research has been supported by the European Regional Development Fund within the joint project of SIA TILDE and University of Latvia “Multilingual Artificial Intelligence Based Human Computer Interaction” No. 1.1.1.1/18/A/148.