Pretraining and Fine-Tuning Strategies for Sentiment Analysis of Latvian Tweets

Gaurish Thakkar[1] and Mārcis Pinnis[2]

[1]University of Zagreb, [2]TILDE

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 Skadiņš
- ⋆ *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 time-frame 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.

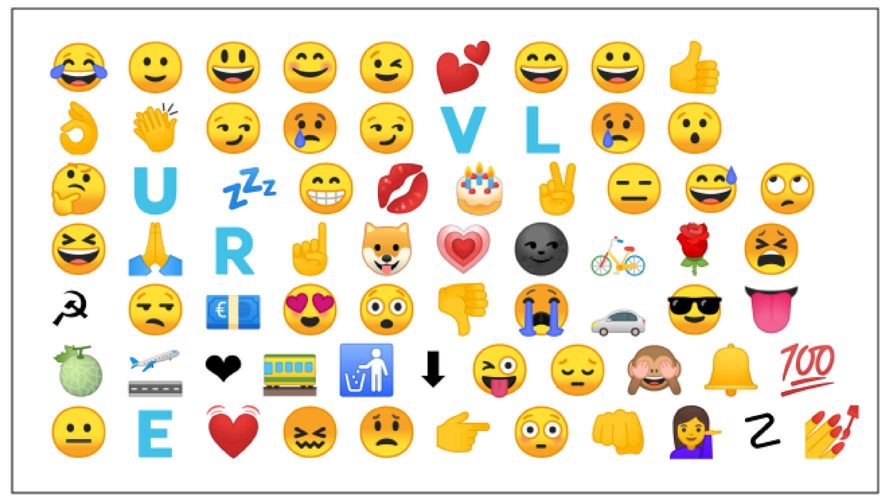

Figure 1:Examples of (non-exhaustive) list of added emoticons

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

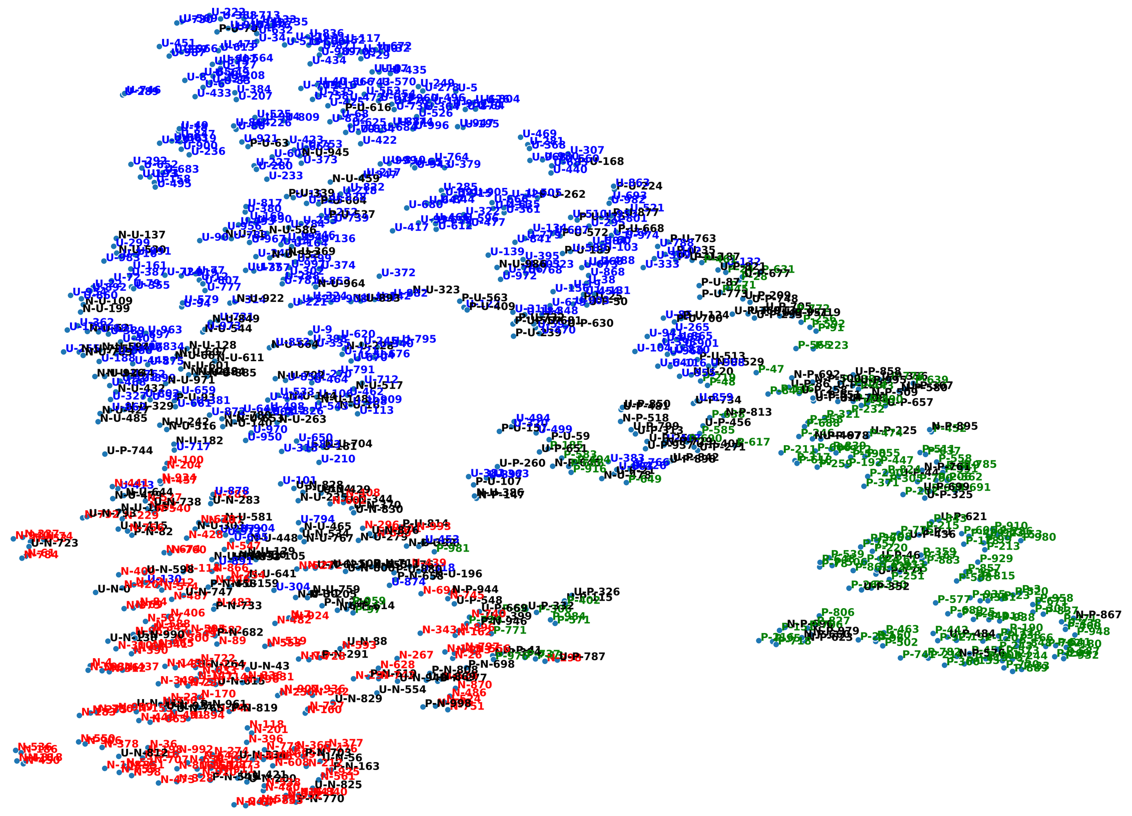

Figure 2:Tweet representation and prediction scatter plot

Results

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

Dataset	Perceptron [1]	mBERT			ALBERT	ELECTRA
		Base	Pre	Pre+Emo		
Gold	0.661	0.678	**0.756**	0.754	0.661	0.711
Gold+Peisenieks	0.676	0.692	0.747	**0.764**	0.698	0.706
Gold+Auto (with ☺)	0.624	0.679	**0.769**	0.748	0.649	0.680
Gold+Auto (no ☺)	0.512	0.523	0.648	**0.660**	0.483	0.621
Gold+Auto (both)	0.487	0.526	0.618	**0.657**	0.509	0.564
Gold+English	0.613	0.698	0.692	**0.720**	0.669	0.684

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.

Get Your Code Here

https://github.com/thak123/bert-twitter-sentiment

Acknowledgements

The work presented in this paper has received funding from the European Union's Horizon 2020 research and innovation programme under the *Marie Sklodowska-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.

References

[1] Pinnis, M. (2018). Latvian Tweet Corpus and Investigation of SentimentAnalysis for Latvian. In Proceedings of Baltic HLT 2018, pages 112–119, Tartu, Estonia. IOS Press.

Contact Information

Email: gthakkar@m.ffzg.hr