Contrastive Learning for Sentiment Classification - Appendix

Studer Justin, Ghodkih Hrishikesh, Neuner-Jehle Joel, Group: Taskforce

Department of Computer Science, ETH Zurich

A. Ablations for other models

To ensure a fair comparison, we conducted ablations on the other models we compared the contrastive approach to. For BCE BERT and GRUBERT, we examined the effect of the learning rate, and for GRUBERT we tried to find the best number of GRUs. We find that our implementation produces the best result at the same GRU configuration as was proposed by (Horne et al., 2020), namely "shared-3".

Learning Rate	1e-4	5e-5	1e-5	5e-6	1e-6
BCE BERT	87.03 (0.15)	87.68 (0.04)	**87.96 (0.07)**	87.89 (0.20)	87.57 (0.37)
GRUBERT	87.17 (0.02)	88.00 (0.23)	**88.05 (0.12)**	87.49 (0.35)	87.26 (0.27)

Table 1. Learning Rate Experiments for the binary classification BERT baseline which was trained using binary cross-entropy and GRUBERT (using the optimal configuration from (Horne et al., 2020) as a starting point, namely shared-3)

Number of GRUs	1	2	3	4	6
GRUBERT	87.52 (0.25)	86.76 (0.18)	**88.05 (0.12)**	87.51 (0.36)	87.70 (0.04)

Table 2. Experiment results when varying the number of GRUs used in GRUBERT. Notice that the value for 3 GRUs stems from the same experiments as the one from Table 3

Finally, we also experimented with more traditional models and used sklearn and nltk library for our experiments. The following tables show comparisons between different configurations and preprocessing methods of these models. First, we determined the best classifier using cross validation on the mean accuracy, whereby all other preprocessing parameters were set to default. We then tested the best model with various preprocessing methods in order to reduce different forms of a single word to its original stem, hence avoiding redundant features. Realizing that these methods did not improve the model, we tweaked the number of n-grams and features independently which fairly improved the classifier model. We put together all the best parameter values and train the classifier one last time to obtain an accuracy score of 82.32 as shown in Table 1 of the main document.

Classifier	Linear SVC	Logistic Regression	Bernoulli NB	Gaussian NB
Accuracy	80.64	**80.84**	74.19	65.59

Table 3. Model selection results using 10-fold cross validation on mean accuracy.

	Preprocessing			
	stopwords	stemming	lemmatizing	score
LogReg				**80.84**
	x			79.16
		x		80.55
			x	80.75
	x	x	x	79.29

Table 4. The effect on the accuracy by performing per-token preprocessing of the input data.

N-grams	1 (Default)	2	3
Accuracy	80.84	**81.99**	81.87

Table 5. Number of n-grams taken into consideration to create the embeddings.

Features	5000	1000 (Default)	15000
Accuracy	80.42	80.84	**80.98**

Table 6. Maximum number of features extracted from the entire dataset and thus also the length of the embedding vector.

References

Horne, L., Matti, M., Pourjafar, P., and Wang, Z. GRUBERT: A GRU-based method to fuse BERT hidden layers for Twitter sentiment analysis. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop*, pp. 130–138, Suzhou, China, December 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.aacl-srw.19.