A. Appendix for SentiCSE

A.1. Hyper-parameter Settings

Tables 1 and 2 show the values of hyper-parameters used in the pre-training of SentiCSE and in the few-shot setting, respectively. For the linear probing setting, we borrowed the configurations of SentEval (Conneau and Kiela, 2018). More details can be found in our code uploaded to the submission system.

Parameters	Values
Max sequence length	128
Batch size	64
Learning rate	1e-5
Number of steps	20,000
Pooler type	cls
Temperature	0.05
λ_w	0.15
Hard negative weight	1
Masking ratio	0.1

Table 1: Hyper-parameters for pre-training of SentiCSE.

Parameters	Values
Batch size	16
Learning rate	1e-5
Number of epochs	100
Evaluation steps	20
Weight decay	0.01

Table 2: Hyper-parameters for the few-shot setting.

A.2. More Analysis on SentiCSE

In the main manuscript, we report the results obtained when MR data is adopted in SentiCSE as a source domain. Here, we tested other data as the source domain of SentiCSE. Table 3 reports the performance obtained from each case on the linear probing, respectively. We observed that the average performance was higher in the order when the source domain was MR, SST-2, Yelp-2, and IMDB. This order is very interesting because, as shown in Table 1 in the main manuscript, it has the same order with the '% of SentiWords' of each dataset. This implies the importance of the amount of sentiment information contained in each sentence when choosing a source domain for pre-training.

Furthermore, we also implemented BERT-based SentiCSE and compared it with the BERT-based baselines: SentiBERT and SentiX. We also compared BERT itself and BERT-based SimCSE with ours. Overall, a similar trend is observed when we compared the RoBERTa-based models including SentiLARE.

Source\Target	IMDB	SST-2	Yelp-2	MR	Avg.
IMDB	94.21^*	89.22	95.79	86.49	91.43
SST-2	90.62	94.72^*	94.86	**89.87**	92.52
Yelp-2	93.67	87.50	97.93^*	86.59	91.42
MR	**94.28**	**95.30**	**96.27**	89.02^*	**93.72**

Table 3: Linear probing performance of SentiCSE on each source-target data combination. For each data column, the accuracy with * indicates the result from a model that was pre-trained on the same dataset as its source domain.

Next, we examined the impact of α used in the sentence-level objective on the performance. Table 4 shows the linear probing results with regards to different α values.

α	IMDB	SST-2	Yelp-2	MR	Avg.
0	94.02	**95.30**	96.30	**89.12**	93.69
1	**94.06**	94.95	**96.27**	88.93	93.55
2	93.99	95.07	**96.27**	88.18	93.38

Table 4: Performance depending on different α.

A.3. More Analysis on SgTS

As mentioned in the main manuscript, we observed the high correlation between our SgTS metric and the few-shot performance. We plot the SgTS scores and corresponding few-shot accuracy as shown in Figure 1 (the Spearman correlation $\rho = 0.96$).

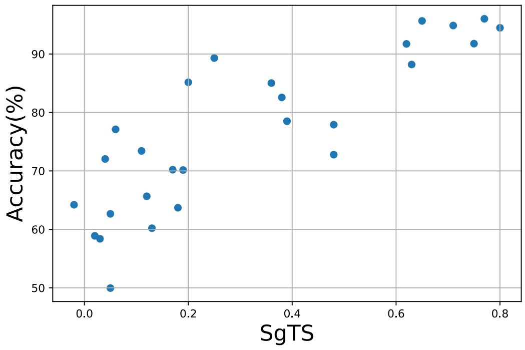

Figure 1: Correlation between the SgTS score and the few-shot accuracy.

Next, we examined the behavior of the SgTS scores compared to the conventional STS as the pre-training of SentiCSE progresses. Figure 2 shows the results. Each score was obtained at every 500 step. We observed that the SgTS score reached to a high value in the very initial stage, which indicates the effectiveness of our pre-training objectives. We also observed that in the first half steps, SgTS gradually increased while STS gradually decreased. This implies the difference between

Loss	Dataset	RoBERTa	SimCSE	SentiBERT	SentiX	SentiLARE	SentiWSP	SentiCSE
Alignment	IMDB	0.01	0.24	0.32	1.12	0.21	0.23	1.37
	SST-2	0.00	0.46	0.22	1.13	0.09	0.32	1.65
	Yelp-2	0.01	0.38	0.30	1.23	0.43	0.23	1.34
	MR	0.00	0.45	0.23	1.11	0.09	0.30	1.61
Uniformity	IMDB	-0.01	-0.48	-0.59	-1.94	-0.38	-0.43	-1.07
	SST-2	-0.01	-0.91	-0.41	-2.01	-0.19	-0.60	-1.02
	Yelp-2	-0.01	-0.72	-0.56	-1.89	-0.71	-0.43	-1.06
	MR	-0.01	-0.89	-0.44	-2.00	-0.18	-0.58	-1.19

Table 5: Representation quality measured by Alignment and Uniformity. The lower the scores, the better the quality.

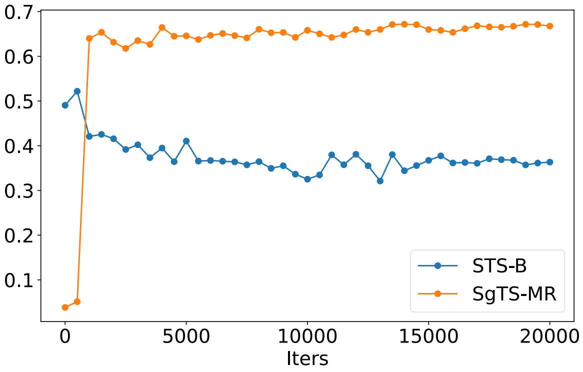

Figure 2: Behavior of the SgTS and STS scores in pre-training.

sentiment-favorable representation and semantic representation. However, after convergence of the two scores, we observed that the STS score converges around 0.4, which demonstrates that understanding semantic information is still helpful in constructing high-quality sentiment representations.

A.4. Alignment and Uniformity Metrics

In terms of evaluation methods for representation quality, Alignment and Uniformity have been known as key properties in contrastive learning (Wang and Isola, 2020). The two factors take alignment between the pairs in the same class and uniformity of the representation space (Gao et al., 2021), respectively. They can measure the quality of learned embeddings. Table 5 shows their scores measured on the constructed representations obtained from each model. It was difficult to find high correlation between the quality of sentiment representations visualized via PCA and the Alignment and Uniformity scores. In our future work, we plan to develop a novel evaluation metric for sentiment representations based on Alignment and Uniformity.

B. References

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6894–6910.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In International Conference on Machine Learning, pages 9929–9939. PMLR.