Revisiting Supervised Contrastive Learning for Microblog Classification

Anonymous ACL submission

Abstract

Microblog content (e.g., Tweets) is noisy due to its informal use of language and its lack of contextual information within each post. To tackle these challenges, state-of-the-art microblog classification models rely on pre-training language models (LMs). However, pre-training dedicated LMs is resource-intensive and not suitable for small labs. Supervised contrastive learning (SCL) has shown its effectiveness with small, available resources. In this work, we examine the effectiveness of fine-tuning transformer-based language models, regularized with a SCL loss for English microblog classification. Despite its simplicity, the evaluation on two English microblog classification benchmarks (TweetEval and Tweet Topic Classification) shows an improvement over baseline models. The result shows that, across all subtasks, our proposed method has a performance gain of up to 11.9 percentage points. All our models are open source.

1 Introduction

Microblog classification is a text classification task on microblog content (e.g., Tweets). State-of-the-art microblog classification models rely on pre-training domain-specific transformer-based language models (LMs), such as Bertweet (Nguyen et al., 2020), XLM-T (Barbieri et al., 2022) and TimeLMs (Loureiro et al., 2022). In comparison, large language models (LLMs) such as ChatGPT and GPT-4 fall short of this task (Kocon et al., 2023). However, pre-training LMs requires large computational resources, which is not feasible for small labs. An affordable alternative is to fine-tune a base pre-trained LM, such as RoBERTa (Liu et al., 2019). In this work, we focus on the fine-tuning approach.

Typically, microblog content is noisy. First, the informal use of language introduces a large volume of incorrect grammar or typos. Second, social

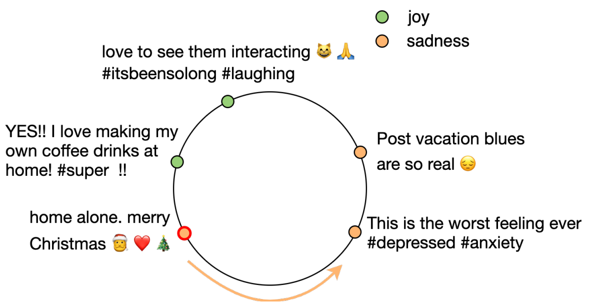

Figure 1: An example of how supervised contrastive learning utilizes label information to form better representation on a hyper-sphere. The orange circle with the red edge represents an ambiguous sentence whose representation can be improved with SCL.

media posts are mostly short in length. Due to the character limit, microblog content often lacks contextual information (Kim et al., 2014), which inherently increases the difficulty for the model to learn a good representation of the data. We hence investigate the use of supervised contrastive learning (SCL) (Khosla et al., 2020; Gunel et al., 2021) for microblog classification.

We suggest that SCL helps improve the learnt representation of models and performance on microblog classification tasks. This is because SCL utilizes label information to enhance the intra-class concentration of features (Saunshi et al., 2019). Figure 1 depicts a common phenomenon in microblog classification, where the model fails to represent an ambiguous sentence (circle with the red edge) in the embedding space. Models trained with a SCL loss explicitly pull the ambiguous sentence closer to the region where semantically similar sentences are located. Therefore features of the same label are more concentrated in the embedding space. The orange arrow represents the "pulling" effect of SCL's learning objective.

Overall, our contributions are:

1. We examine the effectiveness of SCL loss in a

supervised learning setting in terms of downstream performance on two microblog classification tasks, namely, TweetEval[1] (Barbieri et al., 2020) and Tweet Topic Classification[2] (Antypas et al., 2022).

2. We open-sourced a generic fine-tuning framework with SCL (`https://anonymous.4open.science/r/74D1`).

2 Related Work

We provide two lines of literature that are related to our work: microblog classification and contrastive learning in NLP.

2.1 Microblog classification

State-of-the-art models for microblog classification follow the pre-training and fine-tuning supervised learning schema. Pre-trained LMs such as Bertweet (Nguyen et al., 2020) or TimeLMs (Loureiro et al., 2022) provides a good instantiation of model parameters, which often leads to superior performance after fine-tuning on dedicated downstream tasks, such as part-of-speech tagging (Gimpel et al., 2011; Liu et al., 2018; Ritter et al., 2011), named-entity recognition (Strauss et al., 2016) and microblog classification (Barbieri et al., 2020; Rosenthal et al., 2019; Hee et al., 2018). However, pre-training on large scale corpora is not accessible to small labs. Therefore, we focus on the fine-tuning stage with a base LM (RoBERTa), to achieve comparable performance of pre-trained models.

2.2 Contrastive learning in NLP

Two often used contrastive learning algorithms in NLP are self-supervised contrastive learning (SSCL) and SCL. SSCL algorithms such as SimCLR (Chen et al., 2020) learn representations in an instance discrimination task, which is an extreme case of a multi-class classification task, where each instance has its own class. During training, SSCL loss forces a higher inner product of representations between positive pairs than negative pairs. Since SSCL does not require label information, it is ideal for learning sentence-level embeddings (Gao et al., 2021; Wu et al., 2020).

However, learning can be error-prone without label information. This is reflected in the defect of the instance discrimination objective (Wang and Liu, 2021). The pushing apart of negative samples ignores their underlying relations, which causes the breakdown of the formation of certain useful features. Saunshi et al. (2019) provided a theoretical analysis of how negative classes can overlap in the latent space in SSCL, known as class collision.

To account for this problem, SCL leverages label information to enforce a different representation of inherently "similar" samples. Previous work applied SCL loss in NLP for few-shot text classification (Gunel et al., 2021) and showed its effectiveness under the problem of data scarcity. It is evaluated on the GLUE benchmark, which is a collection of nine sentence- or sentence-pair language understanding tasks in the domain of movie reviews and news. Differentiating from their work, we investigate whether SCL is beneficial for regular supervised learning with many labeled data in the domain of microblog classification.

3 Method

To examine the effectiveness of SCL for microblog classification, we train a transformer-based sequence classifier in a supervised learning setting. The learning objective is to minimize a linear combination of a SCL loss and a CE loss.

3.1 Architecture

Given a single-label multi-class text classification dataset χ and a batch size of N_{bs}, a feature extractor $f_\theta(\cdot)$ maps the input sentence, x_n, into two augmented feature vectors $r_i, r_j \in \mathbb{R}^{N_{feature}}$. $N_{feature}$ is the output dimensionality of the feature extractor (768 in our case). Consistent with the original SCL paper (Khosla et al., 2020), the augmented feature vectors are then L2-normalized and fed into a projection network to create the latent representation $h_n = g_\phi(r_n) \in \mathbb{R}^{N_{proj}}$, where the distance matrix is computed. Since this is a sequence classification task, N_{proj} equals the number of classes in the dataset. Cosine similarity is used as the distance measure. In this work, we use the huggingface implementation of *RoBERTa-base*[3] as the feature extractor and a linear layer as the projection network. A detailed architecture diagram is illustrated in Figure 2.

[1] `https://huggingface.co/datasets/tweet_eval`

[2] `https://huggingface.co/cardiffnlp/tweet-topic-19-single`

[3] `https://huggingface.co/roberta-base`

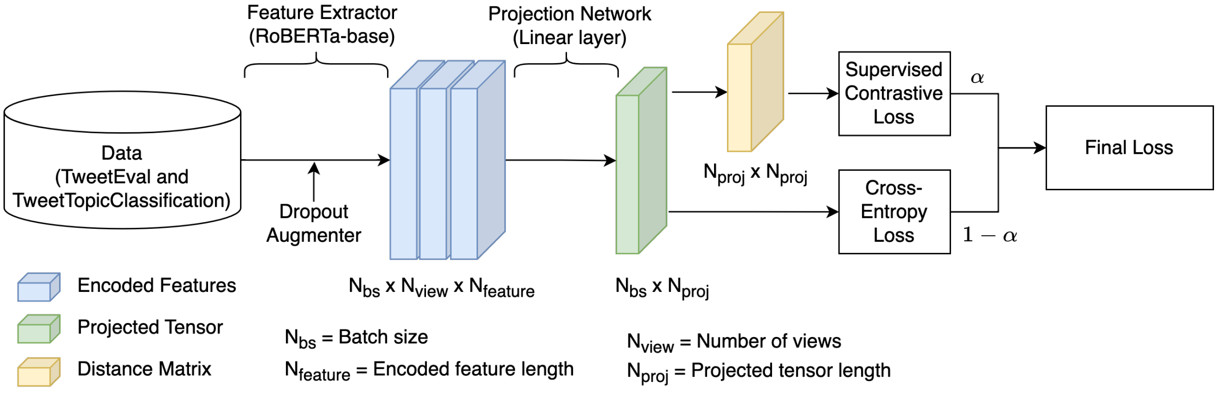

Figure 2: Architecture of the proposed method.

3.2 Losses

Given a multi-view batch of augmented samples with index $i \in I \equiv \{1, 2, ..., 2N_{bs}\}$, the positive pairs are constructed from the augmented views of the same instance, and all other augmented instances with the same label as the anchor. Negative samples are all other augmented instances with different labels from the same batch. Let $P(i)$ and $K(i)$ (with cardinality $|P(i)|$ and $|K(i)|$) be a set of positive and negative samples with index i.

The SCL loss is defined as,

$$\mathcal{L}_{SCL} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{j \in P(i)} \log \frac{\exp(\frac{h_i \cdot h_j}{\tau})}{\sum_{k \in K(i)} \exp(\frac{h_i \cdot h_k}{\tau})} \tag{1}$$

, where $\tau \in \mathbb{R}^+$ denotes the temperature parameter. Note that the summation over $P(i)$ indicates that the SCL loss allows an arbitrary number of positive pairs. The final loss is a linear combination of supervised contrastive loss and a standard CE loss,

$$\mathcal{L}_{final} = \alpha \mathcal{L}_{SCL} + (1 - \alpha)\mathcal{L}_{CE} \tag{2}$$

with a coefficient $\alpha \in [0, 1]$.

4 Evaluation

4.1 Benchmarks

Our method is evaluated on two tweets classification benchmarks, TweetEval (Barbieri et al., 2020) and Tweet Topic Classification (Antypas et al., 2022). In total, eight subtasks are used for evaluation, where seven of which are from TweetEval and one subtask from Tweet Topic Classification.

TweetEval. TweetEval is a benchmark consisting of seven microblog classification subtasks, including *emoji prediction*, *emotion recognition*, *irony detection*, *hate speech detection*, *offensive language identification*, *sentiment analysis* and *stance detection*. Each subtask is collected from the SemEval shared task series from 2016 to 2019.

Tweet Topic Classification. Tweet Topic Classification is a microblog classification benchmark with multi-label and single-label settings. We consider only the single-label setting in our experiment. Six classes are included in this dataset, namely, *arts&culture*, *business&entrepreneurs*, *pop culture*, *daily life*, *sports&gaming* and *science&technology*. Additionally, since the original dataset does not have a validation set, we split 10% of the training set into a validation set.

Preprocessing. A minimal preprocessing step is used in this work. All user mentions are replaced with a "`@user`" special token and links with a "`http`" special token. The masking of user mentions prevents the leaking of real user information.

4.2 Metrics

We use the same evaluation metrics from the original benchmarks. Specifically, for TweetEval, we use macro averaged F1 over all classes, in most cases. There are three exceptions: stance detection (macro-averaged of F1 of favor and against classes[4]), irony detection (F1 of ironic class[5]), and sentiment analysis (macro-averaged recall). A global metric (TE) based on the average of all dataset-specific metrics is as well included. For

[4]Stance detection is a classification task with three labels, namely, favor, against and none.

[5]Irony detection is a binary classification task with two labels, namely, irony and non-irony.

Model	Emoji	Emotion	Hate	Irony	Offensive	Sentiment	Stance	All
ChatGPT^{llm}	18.2	-	-	-	-	63.7	56.4	-
Rob-rt ^{pt}	31.4	78.5	52.3	59.7	77.1	69.1	66.7	61.0
Rob-tw ^{pt}	29.3	72.0	46.9	65.4	80.5	72.6	69.3	65.2
XLM-r^{pt}	28.6	72.3	44.4	57.4	75.7	68.6	65.4	57.6
XLM-tw^{pt}	30.9	77.0	50.8	69.9	79.9	72.3	67.1	64.4
Bertweet^{pt}	33.4	79.3	56.4	**82.1**	79.5	73.4	71.2	**67.9**
TimeLM-19^{pt}	33.4	**81.0**	**58.1**	48.0	**82.4**	73.2	70.7	63.8
TimeLM-21^{pt}	**34.0**	80.2	55.1	64.5	82.2	**73.7**	**72.9**	66.2
Rob-bs (CE)^{ft}	30.9	76.1	46.6	61.7	79.5	71.3	68.0	61.3
Rob-bs (CE+SCL)^{ft}	**32.0**	**78.1**	**49.4**	**68.0**	**79.6**	**72.0**	**69.4**	**64.1**
Metric	M-F1	M-F1	M-F1	F$^{(i)}$	M-F1	M-Rec	AVG(F)	TE

Table 1: Results on TweetEval. We divide three types of models for a fair comparison, namely, pre-trained LMs, LLMs and fine-tuned LMs. Note that our proposed models are fine-tuned RoBERTa-base. Results from pre-trained LMs and LLMs are provided as a reference to evaluate our fine-tuned models. SotA models are bold for each subtasks in each model class indicated by the superscript (llm, pt and ft).

Tweet Topic Classification, we report macro average precision, recall, F1, and accuracy.

4.3 Result

We compare models fine-tuned with a combined SCL and CE loss, compared with models fine-tuned with only CE loss. The choice of hyper-parameters is presented in A.1. All experiments are run with a single NVIDIA RTX A6000 48 GB graphics card, and are run three times with different seeds (0, 1 and 2). Numbers shown in the following section represent the average value over three seeds.

We provide three categories of baseline models, including (a) LLMs (Kocon et al., 2023), (b) pre-trained LMs (Barbieri et al., 2022; Nguyen et al., 2020; Loureiro et al., 2022; Barbieri et al., 2020) and (c) fine-tuned LMs (Barbieri et al., 2020).

TweetEval. We compare *RoBERTa-base* fine-tuned with and without SCL loss in the TweetEval benchmark. All hyper-parameters are shared across seven sub-tasks. We observed (Table 1) that models fine-tuned with the linear combination of a SCL and a CE loss show an improvement, ranging from 0.1 to 8.3 percentage points. Although the performance of our fine-tuned model (CE+SCL) is not as good as the SotA pre-trained LMs, it surpasses the performance by ChatGPT in all subtasks and by its pretrained counterparts in various subtasks.

Tweet Topic Classification. According to results shown in Table 2, the SCL+CE model outperforms the CE baseline on the Tweet Topic Classification benchmark by large margins. Tweet Topic Classification is a single-label classification task with six classes. Moreover, it surpasses the state-of-the-art model presented in the original paper (Antypas et al., 2022).

Model	P	R	F1	Acc
Rob-bs (CE)	64.8	66.7	65.6	85.9
Rob-bs (CE+SCL)	**76.9**	**75.7**	**76.2**	**88.2**
SotA	76.5	68.9	70.0	86.4

Table 2: Results on Tweet Topic Classification. SotA refers to TimeLM-19 (Loureiro et al., 2022).

5 Conclusion

With the observation that user-generated microblog content contains a large volume of noise that is inherent in the dataset, we develop a generic yet simple microblog classification fine-tuning framework with a SCL-based regularizer in the training objective. Our framework improves the baseline variant that is fine-tuned with only a cross-entropy loss by large margins across all tasks on the TweetEval and Tweet Topic Classification benchmarks. On Tweet Topic Classification, our model also surpassed the state-of-the-art models which are pre-trained on microblog-related corpora. The ablation study in Appendix A.2 in shows the importance of utilizing label information for the SCL regularizer. By qualitatively evaluating the model's prediction, we have identified two types of commonly made errors in Appendix A.3.

Acknowledgements

Omitted for double-blind review.

Limitations

Albeit evidence has shown that our training framework improves transformer-based models' performance on English microblog classification tasks. There exist three limitations that we are aware of.

First, other variants of text augmentation techniques have not been experimented with in this work. Contrastive learning as a learning framework learns good representation in terms of good class separability. A critical component that influences learning is data augmentation. Notably, how to do data augmentation on text is by itself an important and challenging topic. We ground our hypothesis based on observations made by others, which use the dropout mechanism in the transformer-based feature extractors. Yet, it is not clear why and how relying on such a simple mechanism creates good results in terms of quality.

Second, microblog classification benchmarks of languages other than English have not been experimented with. Tested on all publically available English microblog classification datasets, we claim that our framework is generic only to English corpora. However, it is interesting to investigate whether it generalizes to other languages as well, in particular, low-resource languages. Yet that adds another layer of complexity, which is learning with limited label information.

Third, the effect of batch size is not experimented with due to the limit in our computational resources. Large batch size is another key hyperparameter that leads to the success of contrastive learning. The upper threshold that is constrained by our GPU device is 96. This includes an anchor batch of size 32 together with its two augmented batches.

Ethics Statement

To our knowledge, this work does not concern any substantial ethical issue. Corpora used in this work are preprocessed by masking all user mentions and links. Example sentences shown in this paper do not harm any individuals or groups. Of course, the application of classification algorithms could always play a role in Dual-Use scenarios. However, we consider our work as not-risk-increasing.

References

Dimosthenis Antypas, Asahi Ushio, José Camacho-Collados, Vítor Silva, Leonardo Neves, and Francesco Barbieri. 2022. Twitter topic classification. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 3386–3400. International Committee on Computational Linguistics.

Francesco Barbieri, Luis Espinosa Anke, and José Camacho-Collados. 2022. XLM-T: multilingual language models in twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 258–266. European Language Resources Association.

Francesco Barbieri, José Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 1644–1650. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

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, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6894–6910. Association for Computational Linguistics.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA - Short Papers*, pages 42–47. The Association for Computer Linguistics.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2021. Supervised contrastive learning for pre-trained language model fine-tuning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. Semeval-2018 task 3: Irony detection in english tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2018, New Orleans, Louisiana, USA, June 5-6, 2018*, pages 39–50. Association for Computational Linguistics.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Suin Kim, Ingmar Weber, Li Wei, and Alice Oh. 2014. Sociolinguistic analysis of twitter in multilingual societies. In *25th ACM Conference on Hypertext and Social Media, HT '14, Santiago, Chile, September 1-4, 2014*, pages 243–248. ACM.

Jan Kocon, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydlo, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, Anna Kocon, Bartlomiej Koptyra, Wiktoria Mieleszczenko-Kowszewicz, Piotr Milkowski, Marcin Oleksy, Maciej Piasecki, Lukasz Radlinski, Konrad Wojtasik, Stanislaw Wozniak, and Przemyslaw Kazienko. 2023. Chatgpt: Jack of all trades, master of none. *CoRR*, abs/2302.10724.

Yijia Liu, Yi Zhu, Wanxiang Che, Bing Qin, Nathan Schneider, and Noah A. Smith. 2018. Parsing tweets into universal dependencies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 965–975. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and José Camacho-Collados. 2022. Timelms: Diachronic language models from twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, ACL 2022 - System Demonstrations, Dublin, Ireland, May 22-27, 2022*, pages 251–260. Association for Computational Linguistics.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 9–14. Association for Computational Linguistics.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1524–1534. ACL.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2019. Semeval-2017 task 4: Sentiment analysis in twitter. *CoRR*, abs/1912.00741.

Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. 2019. A theoretical analysis of contrastive unsupervised representation learning. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5628–5637. PMLR.

Benjamin Strauss, Bethany Toma, Alan Ritter, Marie-Catherine de Marneffe, and Wei Xu. 2016. Results of the WNUT16 named entity recognition shared task. In *Proceedings of the 2nd Workshop on Noisy User-generated Text, NUT@COLING 2016, Osaka, Japan, December 11, 2016*, pages 138–144. The COLING 2016 Organizing Committee.

Feng Wang and Huaping Liu. 2021. Understanding the behaviour of contrastive loss. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 2495–2504. Computer Vision Foundation / IEEE.

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020. CLEAR: contrastive learning for sentence representation. *CoRR*, abs/2012.15466.

A Appendix

A.1 Hyper-parameters

For any anchor sentence, two augmented views are generated via the dropout augmenter. The dropout rate of both the self-attention and linear layer in the transformer-based feature extractor is set to 0.1. We use Adam optimizer with a learning rate of $1e-5$. The learning rate is warmed up for 10 epochs. Warming up the learning rate at the beginning of the training phase prevents the model from early over-fitting. The total number of training epochs varies for all tasks, since we use early stopping on the validation set with a patience of 5 epochs. We conduct a hyper-parameter search on the SCL loss ratio $\alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and the temperature parameter $\tau \in \{0.03, 0.1, 0.3, 0.5, 0.7, 0.9\}$. The best combination is $\alpha = 0.5$ and $\tau = 0.9$. Note that we use a batch size of 32, so the augmented batch contains 96 instances. This is extremely small compared with other work in contrastive learning, which suggests larger batch size benefits learning. However, due to the upper limit of the GPU used in our lab, we can not conduct experiments investigating the effect of a larger batch size.

Model	Emoji	Emotion	Hate	Irony	Offensive	Sentiment	Stance	All
Rob-bs (CE+SCL)	**32.0**	**78.1**	**49.4**	**68.0**	**79.6**	**72.0**	**69.4**	**64.1**
Rob-bs (CE+SSCL)	25.3	59.4	40.2	55.2	79.4	71.8	60.6	56.0
Metric	M-F1	M-F1	M-F1	$F^{(i)}$	M-F1	M-Rec	AVG(F)	TE

Table 3: Results on models fine-tuned with a SSCL and a CE loss, compared with the same model fine-tuned with a SCL and a CE loss, evaluated on TweetEval.

(a) CE+SCL

(b) CE+SSCL

Figure 3: Confusion matrix on the emotion detection subtask.

Model	Pr	Recall	F1	Acc
Rob-bs (CE+SCL)	**74.3**	**76.0**	**74.9**	**88.2**
Rob-bs (CE+SSCL)	63.4	57.4	43.5	33.0

Table 4: Ablation study result on models fine-tuned with SSCL loss and CE loss, compared with the same model fine-tuned with SCL loss and CE loss, evaluated on Tweet Topic Classification.

A.2 Ablation Study

To remove the effect of SCL's intrinsic negative mining property, We conducted an ablation study on replacing the SCL loss term with a SSCL loss term, while keeping the CE loss. The motivation is to study the importance of label information in learning the representation of microblog texts. The model is evaluated on the same benchmarks above.

Quantitative experiments. Experiment details including architecture and evaluation in the SSCL setting are identical to all other experiments, as described in Section 3.1 and Section 4. SSCL is an instance discrimination task with the following loss in Equation 3.

$$\mathcal{L}_{SCCL} = -\log \frac{\exp(h_i \cdot h_j/\tau)}{\sum\limits_{k \in K(i)} \exp(h_i \cdot h_k/\tau)} \quad (3)$$

The implementation difference is only shown in the computation of the negative log-likelihood, compared with the SCL loss. Specifically, the SSCL loss does not include a summation over positive pairs of the same label as in Equation 1, as well as the summation over the "true" negative pairs whose labels are different. This indicates that SSCL does not create an averaged representation over all positive samples. Therefore, the pulling and pushing effect of SSCL ignores information carried by distances between other positive samples, leading to a higher chance of creating a worse representation. Being able to consider multiple positives and negatives as in SCL, the model creates more separable features, resulting in a more robust clustering of the representation space.

Table 3 and Table 4 show the result of the classification performance on TweetEval and Tweet Topic Classification, respectively. A noticeable difference in performance, compared with models fine-tuned with SCL and CE, is observed.

Qualitative study. To investigate qualitatively the different behaviors on both classifiers, we first provide the confusion matrices evaluated on the *Emotion Detection* (test set) subtask in TweetEval, as shown in Figure 3. We notice the CE+SSCL model creates 17.3% (44 absolute counts) false pre-

Sentences	SCL	SSCL	True Labels
@user @user Yip. Coz he's a miserable huffy guy 😊	anger	joy	anger
And let the depression take the stage once more 🙃	sadness	joy	sadness
I'm legit in the worst mood ever. #annoyed #irritated	anger	sadness	anger
Of course I've got a horrible cold and am breaking out 2 days before grad 👍 👍 👍 👍	sadness	joy	sadness
the thing about living near campus during the summer is that it's a ghost town but now everyone is back and im #annoyed	anger	sadness	anger
I need a beer #irritated	anger	sadness	anger

Table 5: Ablation study result on models fine-tuned with SSCL loss and CE loss, compared with the same model fine-tuned with SCL loss and CE loss, evaluated on TweetEval.

dictions more than the CE+SCL model. Additionally, we draw samples that are correctly classified in the CE+SCL model while being falsely classified in the CE+SSCL variation in Table 5. Interestingly, 38.6% (39 out of 101) of those samples contain emojis, while 23.3% (330 out of 1421) of the full test set contains emojis. We observe that the use of certain emojis creates ambiguous predictions. It is likely that the model overfits to emojis that lead to misinterpretations. For example, a smiley emoji (😊) does not necessarily entail positive emotions. Utilizing label information, as in SCL, one can enforce the model to avoid over-fitting to such misleading information. Since the scope of this study is not to study noises that the model overfits, we leave this investigation to future work.

A.3 Error Analysis

By inspecting the classification result, we have identified the following two types of texts that are commonly falsely classified by the CE+SCL model.

First, texts that lack contextual cues. Such sentences are either very short, such as *"Duty calls."*; or impossible to the annotators to interpret without further information, such as *"@user @user Can you falter Katli?"* and *"@user Haha nightmare"*. The characteristic of microblog posts inevitably allows for different ways of interpreting the sentences. Thus, it is natural for annotators to embed this uncertainty in the data.

Second, texts whose ground truth label is ambiguous to our evaluation. For example, *"Binge watching #revenge im obsessed."* is labeled as `anger`, while the model's prediction is `joy`. *"Don't grieve over things so badly.."* is labeled as `sadness` and the model's prediction is `optimism`. The annotation process of microblog classification corpora often adopts a generous post-aggregation strategy, leading to the phenomenon where instances with low inter-annotator agreement are not discarded. We acknowledge, that the noise in labels creates another difficulty for any classification model.

To conclude, we realize that the majority of the falsely classified sentences have, to some extent, various levels of ambiguities in the labels. The ambiguities are mainly introduced by the characteristic of microblog posts (e.g., lack of contextual information in microblog posts), or in the annotation process (e.g., a high inclusive rate in the annotation phase).