EICO: Improving Few-Shot Text Classification via Explicit and Implicit Consistency Regularization

Lei Zhao **Cheng Yao**
Alibaba Group
{sashi.zl,cheng.yaoc}@alibaba-inc.com

Abstract

While the prompt-based fine-tuning methods had advanced few-shot natural language understanding tasks, self-training methods are also being explored. This work revisits the consistency regularization in self-training and presents explicit and implicit consistency regularization enhanced language model (EICO). By employing both explicit and implicit consistency regularization, EICO advances the performance of prompt-based few-shot text classification. For implicit consistency regularization, we generate pseudo-label from the weakly-augmented view and predict pseudo-label from the strongly-augmented view. For explicit consistency regularization, we minimize the difference between the prediction of the augmentation view and the prediction of the original view. We conducted extensive experiments on six text classification datasets and found that with sixteen labeled examples, EICO achieves competitive performance compared to existing self-training few-shot learning methods.

1 Introduction

Recently, (Schick and Schütze, 2021a,b) proposed a cloze-style few-shot learning method, PET. By filling the gap between pre-training and fine-tuning with prompt and verbalizer, PET achieved competitive performance with GPT-3 (Brown et al., 2020) with smaller language models like BERT (Devlin et al., 2019), T5 (Lester et al., 2021) and GPT-2 (Radford et al.; Liu et al., 2021). However, the lack of labeled data still limits the performance of few-shot learning. Although acquiring labeled data is costly, unlabeled data is relatively easy to obtain. Leveraging unlabeled data to improve the performance of the few-shot language model through semi-supervised learning is a promising way. In this paper, we focus on advancing the performance of the few-shot language model via semi-supervised learning.

Semi-supervised learning is a method that can leverage both labeled and unlabeled data. A common yet effective semi-supervised method is self-learning, which uses a trained teacher model to generate pseudo-labels for unlabeled examples, and then trains a student model on both labeled and pseudo-labeled examples to utilize the domain-relevant information contained in the unlabeled data. FixMatch (Sohn et al., 2020) and its natural language processing adaption SFLM (Chen et al., 2021) are robust self-training implementations, which generate pseudo-label on the weakly-augmented view of an example and predict pseudo-label by the strongly-augmented view. However, the implicit consistency regularization introduced by the above training scheme may be sub-optimal since class distribution information is lost in pseudo-label generation.

To address the above problem, we propose explicit and implicit consistency regularization enhanced language model (EICO). Beside implicit consistency regularization, EICO utilizes explicit consistency regularization by minimizing the difference between the prediction of the augmented view and the prediction of the original view. To validate the effectiveness and robustness of EICO, we conduct extensive experiments on six natural language understanding tasks. The result of our experiments confirms that EICO can be leveraged to improve the performance of the few-shot text classification. Moreover, we find that EICO presents robustness among different low resources situations and different model size by ablation study.

2 Methodology

2.1 Problem Setting

In this paper, we study the task of learning a model to map an input $x \in \mathcal{X} \subseteq R^d$ onto a label $y \in \mathcal{Y}$. Moreover, in semi-supervised learning, the training dataset consists of labeled examples and unlabeled

Findings of the Association for Computational Linguistics: ACL 2022, pages 3582 - 3587

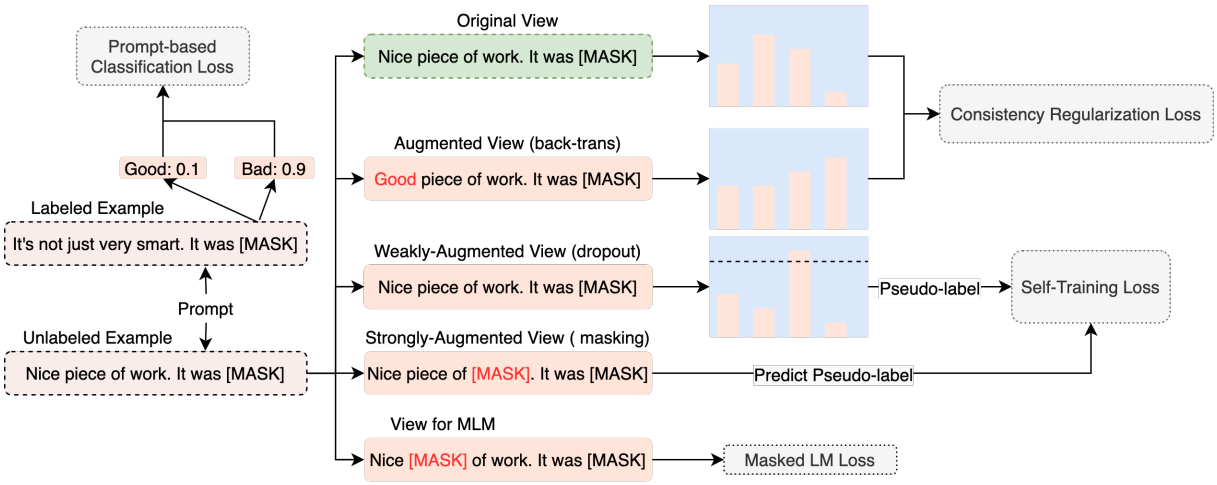

Figure 1: The framework of EICO. Firstly, the appropriate prompt will be added to the input example. Secondly, multiple augmented view will be generated from the prompted input example. Lastly, the corresponding loss function will be applied to the different combinations of views. The augmentation strategies presented in this figure are for demonstration. We will explore multiple combinations of the augmentation strategies in Section 3.

examples. Let $\mathbf{D}_l$ be the labeled training examples, and $\mathbf{D}_u$ be the unlabeled examples, and they are defined as follows:

$$\begin{aligned} \mathbf{D}_l &:= \{(x_i, y_i), i = 1, \ldots, N_l\} \\ \mathbf{D}_u &:= \{x_i^u, i = 1, 2, \ldots, N_u\} \end{aligned} \tag{1}$$

where N_l is the number of the labeled examples, and N_u is the number of the unlabeled examples.

2.2 EICO

In EICO, we leverage a loss function contains prompt-based classification loss, self-training loss, explicit consistency regularization loss, and masked language modeling loss, which will be described in the following sections.

Prompt-based Classification Loss: The prompt-based fine-tuning method was proposed by PET (Schick and Schütze, 2021a,b), where the gap between the natural language classification tasks and the masked language modeling pre-training task are filled by the prompts and verbalizers. The probability of class prediction $y_i \in \mathcal{Y}$ is defined as following:

$$p_m\left(y_i \mid x_i\right) = p_m\left([\text{MASK}] = \mathcal{M}\left(y_i \mid x_i^{\text{prompt}}\right)\right). \tag{2}$$

where $\mathcal{M}$ is a verbalizer from class labels to the corresponding words, and x_i^{prompt} is the reconstructed input sentence with the template.

The pre-trained masked language modeling head can produce the probability over the label word from the verbalizer instead of placing a linear classifier on the top of the backbone model. Therefore, we could use the following cross-entropy loss:

$$\mathcal{L}_{ce} = \frac{1}{B}\sum_{i=1}^{B} H\left(y_i, p_m\left(y_i \mid x_i\right)\right) \tag{3}$$

where B is the batch size. $H(\cdot)$ is the cross-entropy function.

Self-training Loss: After FixMatch (Sohn et al., 2020), SFLM (Chen et al., 2021) demonstrates that the method which generate pseudo-label on the strongly-augmented view and predict pseudo-label on weakly-augmented view can be transferred to natural language understanding tasks.

For each unlabeled example $\mathbf{x}_i^u$, we obtain the weakly-augmented version $\alpha\left(\mathbf{x}_i^u\right)$ and the strongly-augmented version $\mathcal{A}\left(\mathbf{x}_i^u\right)$, where α and $\mathcal{A}$ refers to the augmentation strategies correspondingly. The self-training process consists of two stages. Firstly, we assign a pseudo label to each unlabeled sentence in the batch by computing the output probability distribution corresponding to the weakly-augmented input sentence $\alpha\left(\mathbf{x}_i^u\right)$, defined as $q_i = p_m\left(y_i \mid \alpha\left(\mathbf{x}_i^u\right)\right)$. The pseudo label, $\hat{q}_i$, is obtained by $\hat{q}_i = \arg\max\left(q_i\right)$. Secondly, we compute the prompt-based cross-entropy loss between $\hat{q}_i$ and the prediction corresponding to the strongly-augmented input sentence $\mathcal{A}\left(\mathbf{x}_i^u\right)$. The self-training loss is defined as follows,

$K = 16$	Yelp	AGNews	SST-5	MPQA	SUBJ	QQP	Avg.
$\mathcal{L}_{ce} + \mathcal{L}_{mlm}$ (MLM)	86.74(1.1)	81.63(2.1)	42.37(1.5)	81.68(4.7)	87.92(2.1)	63.45(6.4)	73.96(3.0)
UDA (MLM + $\mathcal{L}_{con}$)							
a^{BT}	89.28(0.9)	83.99(1.0)	42.14(1.2)	80.58(4.3)	87.54(1.8)	62.2(8.9)	74.29(3.0)
a^{MASK}	88.77(1.1)	84.36(0.6)	41.56(2.7)	81.45(3.7)	88.8(1.2)	62.16(9.5)	74.52(3.1)
SFLM (MLM + $\mathcal{L}_{st}$)							
$\alpha^{\text{BT}}, \mathcal{A}^{\text{MASK}}$	88.24(2.2)	84.37(1.1)	41.95(1.5)	80.16(5.5)	89.04(1.3)	61.63(11.2)	74.23(3.8)
$\alpha^{\text{DROP}}, \mathcal{A}^{\text{MASK}}$	88.32(2.1)	84.96(0.5)	42.19(1.3)	81.30(4.5)	88.52(1.2)	61.62(11.1)	74.48(3.5)
Ours							
$a^{\text{BT}}, \alpha^{\text{BT}}, \mathcal{A}^{\text{MASK}}$	89.33(3.0)	84.99(0.6)	42.81(1.7)	81.19(3.8)	89.66(1.3)	63.46(10.8)	75.24(3.5)
$a^{\text{BT}}, \alpha^{\text{DROP}}, \mathcal{A}^{\text{MASK}}$	90.11(0.5)	**85.72**(0.6)	**43.07**(1.8)	81.39(3.5)	88.95(2.4)	63.37(11.0)	75.43(3.3)
$a^{\text{MASK}}, \alpha^{\text{BT}}, \mathcal{A}^{\text{MASK}}$	89.46(1.0)	85.16(0.7)	41.4(3.4)	81.19(3.8)	**90.22**(0.7)	65.67(4.0)	75.52(2.3)
$a^{\text{MASK}}, \alpha^{\text{DROP}}, \mathcal{A}^{\text{MASK}}$	**89.94**(1.4)	84.74(1.1)	41.44(3.4)	**82.30**(3.4)	89.71(1.0)	**66.66**(4.4)	**75.80**(2.5)

Table 1: Main results. We use 16 labeled examples for each class. The results are the average metric of 5 different random seeds. The bold text indicates the best performance on the specific dataset, and the number in brackets is the standard deviation. The last column report the average score over six datasets. a, α and $\mathcal{A}$ indicate different augmentation strategy with superscript. The implementation details are explained in Section 3.

$$\mathcal{L}_{st} = \frac{1}{\mu B} \sum_{i=1}^{\mu B} \mathbb{1}\left(\max\left(q_i\right) \geq \tau\right) H\left(\hat{q}_i, p_m\left(y_i \mid \mathcal{A}\left(x_i^u\right)\right)\right) \tag{4}$$

where $\mathbb{1}(\cdot)$ is an indicator function, τ defines the threshold above which we retain a pseudo-label and μ is the unlabeled example ratio.

Consistency Regularization Loss: EICO adds a explicit consistency regularization loss in the training procedure. Motivated by UDA (Xie et al., 2020) and (Lowell et al., 2021), we minimize the Kullback–Leibler divergence between the augmented view of the example and the original view of example. For the loss function, inspired by R-Drop (Liang et al., 2021), EICO adopts the loss function $\mathcal{L}_{con}$ as follows:

$$\mathcal{L}_{con} = \frac{1}{2}\left(\mathcal{D}_{\text{KL}}\left(p_m(y_i|x_i^u) \| p_m\left(y_i \mid a(x_i^u)\right)\right) + \mathcal{D}_{\text{KL}}\left(p_m\left(y_i \mid a(x_i^u)\right) \| p_m\left(y_i \mid x_i^u\right)\right)\right) \tag{5}$$

where a is an augmentation strategy used in consistency regularization. And $\mathcal{D}_{\text{KL}}$ is the Kullback–Leibler divergence.

Above all, EICO minimize the following loss function:

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda_1 \mathcal{L}_{st} + \lambda_2 \mathcal{L}_{con} + \lambda_3 \mathcal{L}_{mlm} \tag{6}$$

where $\mathcal{L}_{mlm}$ is the masked language modeling loss introduced in BERT (Devlin et al., 2019), λ_1, λ_2, λ_3 are hyper-parameters.

3 Experiment

3.1 Setup

We evaluate our model on six datasets of different natural language understanding tasks including Yelp (Zhang et al., 2015) and SST-5 (Socher et al., 2013) for sentiment analysis, AG's News for news classification, MPQA (Wiebe et al., 2005) for opinion polarity classification, SUBJ (Pang and Lee, 2004) for subjectivity classification, and QQP (Dolan and Brockett, 2005) for semantic equivalence classification.

In semi-supervised learning setting, We set the number of the training labeled examples $K = 16$ for each class (and the number of the development examples is also 16 for each class), the unlabeled data ratio is $\mu = 20$. Following (Gao et al., 2021a), we randomly sample five different splits of $(\mathbf{D}_l^{\text{train}}, \mathbf{D}_l^{\text{dev}}, \mathbf{D}_u)$ from the original training set. Five different models are trained with these splits. Then, we report the average performance of these five models on the original development set. Following (Sohn et al., 2020), We set τ to 0.95. The batch size $B = 16$. The learning rate is set to $1e{-}5$. The sequence length is set to 256. We report main result based on the six-layer pre-trained language model namely DistilRoBERTa-base (Sanh et al., 2019). In our experiment, most of runs achieve best metrics on development set within 150 optimization steps, therefore, we set max optimization steps to 200. For simplicity, we set $\lambda_1, \lambda_2, \lambda_3$ to 1. Following (Perez et al., 2021), we only use development set of K examples to select the best

model, and all hyper-parameters are not tuned with external datasets.

3.2 Baseline

We compare our EICO to several baselines including consistency regularization method, i.e., UDA (Xie et al., 2020) ($L_{ce} + \lambda_2 \mathcal{L}_{con} + \lambda_3 \mathcal{L}_{mlm}$) and self-training method, i.e., SFLM (Chen et al., 2021) ($\mathcal{L}_{ce} + \lambda_1 \mathcal{L}_{st} + \lambda_3 \mathcal{L}_{mlm}$). For exploring the impact of augmentation strategies, we select dropout $(\cdot)^{\text{DROP}}$ (Gao et al., 2021b; Liang et al., 2021), back-translation $(\cdot)^{\text{BT}}$ (Xie et al., 2020) and random token masking $(\cdot)^{\text{MASK}}$ (Devlin et al., 2019) in our experiments. For back-translation augmentation, we use a online translation system publicly available in AliYun[1]. We use French as the middle language. For dropout, we set the embedding dropout rate to 0.1. For random token masking, we randomly mask 15% tokens.

3.3 Main Result

From Table 1 we can see both two types of consistency regularization method outperform the masked language modeling baseline method. Surprisingly, with only explicit consistency regularization (UDA) applied, the random token masking strategy is slightly better than the sophisticated back-translation strategy. And for SFLM method, using dropout as weakly augmentation strategy is also slightly outperform the back-translation strategy. And for ours EICO, the performance of four combinations of the augmentation strategies are both better than SFLM and UDA, which demonstrates that EICO has better ability to leverage the information provided by the unlabeled examples. Within the results from EICO, using random token masking as the augmentation strategy in regularization consistency, and dropout as weak augmentation, random token masking as strong augmentation has the best performance by average, which improves 1.84% compared to baseline and 1.32% compared to SFLM.

3.4 Impact of Model Size

In order to explore the robustness of EICO on different model size, we use a larger language model namely RoBERTa-base (Liu et al., 2019) for ablation study, which have twelve-layer transformer blocks. The hyper-parameters keep the same as Table 1. We report the best performance among all combinations of the augmentation strategies in Table 2. As a result, we found that EICO can achieve competitive performance among two pre-trained language model with different size consistently.

	DistilRoBERTa	RoBERTa
MLM	73.96	77.81
UDA	74.52	77.74
SFLM	74.48	77.65
EICO (Ours)	**75.80**	**78.53**

Table 2: Impact of Model Size. We report the best performance among all combinations of the augmentation strategies. DitilRoBERTa (Sanh et al., 2019) is a six-layer version of RoBERTa, RoBERTa (Liu et al., 2019) is a twelve-layer base model.

3.5 Impact of K

	$K = 8$	$K = 32$
MLM	72.48	76.05
UDA	73.88	76.86
SFLM	73.08	76.44
EICO (Ours)	**74.18**	**77.21**

Table 3: Impact of K. We modified K with all other hyper-parameters keep the same as main result in Table 1. The model is a six-layer DistilRoBERTa.

In order to explore the robustness of EICO on different number of labeled examples K, we conduct experiments on $K = 8$ and $K = 32$, and the rest of hyper-parameters keep the same as in Table 1. From Table 3, we can find that EICO consistently outperform baseline methods in different K.

4 Conclusion

In this work, we propose EICO, a simple yet effective self-training prompt-based few-shot text classification method, where explicit consistency regularization is provided by the agreement of the augmented views of example, and implicit consistency regularization by the pseudo-label technique are utilized. We conducted comprehensive experiments over six text classification datasets and found that EICO outperformed existing methods. Moreover, in the ablation study, we explore the impact of the number of labeled examples K and two different model sizes and found that EICO can consistently achieve competitive performance.

[1] https://www.aliyun.com/product/ai/alimt

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Yiming Chen, Yan Zhang, Chen Zhang, Grandee Lee, Ran Cheng, and Haizhou Li. 2021. Revisiting self-training for few-shot learning of language model. 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 9125–9135. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing, IWP@IJCNLP 2005, Jeju Island, Korea, October 2005, 2005*. Asian Federation of Natural Language Processing.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021a. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3816–3830. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021b. 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.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. 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 3045–3059. Association for Computational Linguistics.

Xiaobo Liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, and Tie-Yan Liu. 2021. R-drop: Regularized dropout for neural networks. *arXiv preprint arXiv:2106.14448*.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021. Gpt understands, too. *arXiv preprint arXiv:2103.10385*.

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. *arXiv preprint arXiv:1907.11692*.

David Lowell, Brian E. Howard, Zachary C. Lipton, and Byron C. Wallace. 2021. Unsupervised data augmentation with naive augmentation and without unlabeled data. 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 4992–5001. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, 21-26 July, 2004, Barcelona, Spain*, pages 271–278. ACL.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. *arXiv preprint arXiv:2105.11447*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Timo Schick and Hinrich Schütze. 2021a. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 255–269. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021b. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 2339–2352. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1631–1642. ACL.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2):165–210.

Qizhe Xie, Zihang Dai, Eduard H. Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 649–657.