

Semi-Supervised Models via Data Augmentation for Classifying Interactive Affective Responses

Jiaao Chen ^{*1}, Yuwei Wu^{*2}, and Diyi Yang¹

¹ Georgia Institute of Technology, Atlanta GA 30318, USA

² Shanghai Jiao Tong University, Shanghai, China

jiaaochen@gatech.edu, will8821@sjtu.edu.cn, diyi.yang@cc.gatech.edu

Abstract. We present semi-supervised models with data augmentation (SMDA), a semi-supervised text classification system to classify interactive affective responses. SMDA utilizes recent transformer-based models to encode each sentence and employs back translation techniques to paraphrase given sentences as augmented data. For labeled sentences, we performed data augmentations to uniform the label distributions and computed supervised loss during training process. For unlabeled sentences, we explored self-training by regarding low-entropy predictions over unlabeled sentences as pseudo labels, assuming high-confidence predictions as labeled data for training. We further introduced consistency regularization as unsupervised loss after data augmentations on unlabeled data, based on the assumption that the model should predict similar class distributions with original unlabeled sentences as input and augmented sentences as input. Via a set of experiments, we demonstrated that our system outperformed baseline models in terms of F1-score and accuracy.

Keywords: Semi-Supervised Learning · Data Augmentation · Deep Learning · Social Support · Self-disclosure

1 Introduction

Affect refers to emotion, sentiment, mood, and attitudes including subjective evaluations, opinions, and speculations [23]. Psychological models of affect have been utilized by other extensive computational research to operationalize and measure users' opinions, intentions, and expressions. Understanding affective responses with in conversations is an important first step for studying affect and has attracted a growing amount of research attention recently [20, 4, 19]. The affective understanding of conversations focuses on the problem of how speakers use emotions to react to a situation and to each other, which can help better understand human behaviors and build better human-computer-interaction systems.

However, modeling affective responses within conversations is relatively challenging since it is hard to quantify the affectiveness [16] and there are no large-scale labeled dataset about affective levels in responses. In order to facilitate

* Equal Contribution. This work was done when Yuwei is a visiting student at Georgia Tech.

research in modeling interactive affective responses, [8] introduced a conversation dataset, OffMyChest, building from Reddit, and proposed two tasks: (1) Semi-supervised learning task: predict labels for Disclosure and Supportiveness in sentences based on a small amount of labeled and large unlabeled training data; (2) Unsupervised task: design new characterizations and insights to model conversation dynamics. The current work focused on the first task.

With limited labeled data and large amount of unlabeled data being given, to alleviate the dependence on labeled data, we combine recent advances in language modeling, semi-supervised learning on text and data augmentations on text to form Semi-Supervised Models via Data Augmentation (SMDA). SMDA consists of two parts: supervised learning over labeled data (Section 4.1) and unsupervised learning over unlabeled data (Section 4.2). Both parts utilize data augmentations to enhance the learning procedures. Our contributions in this work can be summarized into three parts: analysed the OffMyChest dataset in Section 3, proposed a semi-supervised text classification system to classify interactive affective responses classification in Section 4 and described the experimental details and results in Section 5.

2 Related Work

Transformer-based Models : With transformer-based pre-trained models becoming more and more widely-used, pre-training and fine-tuning framework [7] with large pre-trained language models are applied into a lot of NLP applications and achieved state-of-the-art performances [18]. Language models [15, 7, 21] or masked language models [3, 10] are pre-trained over a large amount of text from Wikipedia and then fine-tuned on specific tasks like text classifications. Here we built our SMDA system based on such framework.

Data Augmentation on Text : When the amount of labeled data is limited, one common technique for handling the shortage of data is to augment given data and generate more training “augmented” data. Previous work has utilized simple operations like synonym replacement, random insertion, random swap and random deletion for text data augmentation [17]. Another line of research applied neural models for augmenting text by generating paraphrases via back translations [18] and monotone submodular function maximization [9]. Building on those prior work, we utilized back translations as our augment methods on both labeled and unlabeled sentences.

Semi-Supervised Learning on Text Classification : One alternative to deal with the lack of labeled data is to utilize unlabeled data in the learning process, which is denoted as Semi-Supervised Learning (SSL), since unlabeled data is usually easier to get compared to labeled data. Researchers has made use of variational auto encoders (VAEs) [2, 22, 6], self-training [11, 5, 12], consistency regularization [14, 13, 18] to introduce extra loss functions over unlabeled data to help the learning of labeled samples. VAEs utilize latent variables to reconstruct input labeled

and unlabeled sentences and predict sentence labels with these latent variables; self-training adds unlabeled data with high-confidence predictions as pseudo labeled data during training process and consistency regularization forces model to output consistent predictions after adding adversarial noise or performing data augmentations to input data. We combined self-training, entropy minimization and consistency regularization in our system for unlabeled sentences.

3 Data Analysis and Pre-processing

Researching how human initiate and hold conversations has attracted increasing attention those days, as it can help us better understand how human behave over conversations and build better AI systems like social chatbot to communicate with people. In this section, we took a closer look at the conversation dataset, OffMyChest [8], for better understanding and modeling interactive affective responses. Specifically, we describe certain characteristics of this dataset and our pre-processing steps.

3.1 Label Definition

For each comment of a post on Reddit, [8] annotated them with 6 labels: *Information_disclosure* representing some degree of personal information in comments; *Emotional_disclosure* representing comments containing certain positive or negative emotions; *Support* referring to comments offering social support like advice; *General_support* representing that comments are offering general support through quotes and catch phrases, with *Information_support* offering specific information like practical advice, and *Emotional_support* offering sympathy, caring or encouragement. Each comment can belong to multiple categories.

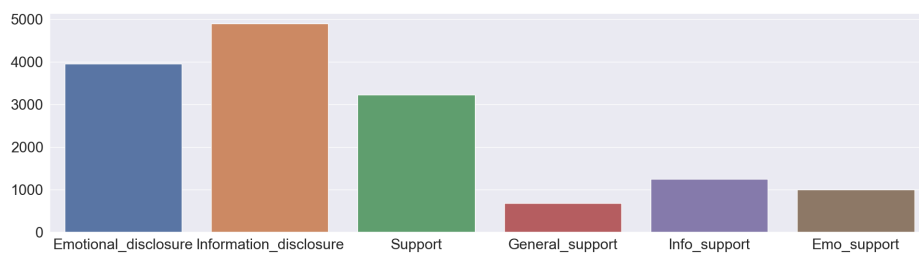


Fig. 1. Distribution of each label in the labeled corpus. The y axis is the number of sentences that have the corresponding labels.

3.2 Data Statics

In OffMyChest corpus, there are 12,860 labeled sentences and over 420k unlabeled sentences for training, 5,000 unlabeled sentences for test. The label distributions of labeled sentences are showed in Fig. 1. To train and evaluate our

Table 1. Dataset split statistics. We utilized both labeled data and unlabeled data for training, generated dev and test set by sampling from given labeled comments set.

Labeled Train Set	Dev Set	Test Set	Unlabeled Train Set
8,000	2,000	2,860	420,607

Table 2. Paraphrase examples generated via back translation from original sentences into augmented sentences.

Original	Augmented	Labels
I'm crying a lot of tears of joy right now.	Right now I'm crying a lot of happy tears.	Emo_disclosure
Stepdad will be the one walking me down the aisle when I get married.	It will be my stepfather walking me down the aisle when I get married.	Info_disclosure
Hope you have a nice day.	I hope you have a good day.	Support
Your best effort, both of you	Both of you are giving it your best shot.	General_support
Plan your transition back to working outside of the home.	Plan your move back to a job outside your own home.	Info_support
I am so freaking happy for you!	I'm so excited for you!	Emo_support

systems, we randomly split the given labeled sentence set into train, development and test set. The data statics are shown in Table 1. We tuned hyper-parameters and chose best models based on performance on dev set, and reported model's performance on test set.

3.3 Pre-processing

We utilized XLNet-cased-based tokenizer³ to split each sentence into tokens. We showed the cumulative sentence length distribution in Fig. 2, 95% comments have less than 64 tokens. Thus we set the maximum sentence length to 64, and remained the first 64 tokens for sentences that exceed the limit. As for data augmentations, we made use of back translation with German as middle language to generate paraphrases for given sentences. Specifically, we loaded translation model from Fairseq⁴, translated given sentences from English to German, and then translated them back to English. Also to increase the diversity of generated paraphrases, we employed random sampling with a tunable temperature (0.8) instead of beam search for the generation. We describe some examples in Table 2.

4 Method

We convert this 6-class affective response classification task into 6 binary classification tasks, namely whether each sentence belongs to each category or not

³ https://huggingface.co/transformers/model_doc/xlnet.html#xlnettokenizer

⁴ <https://github.com/pytorch/fairseq>

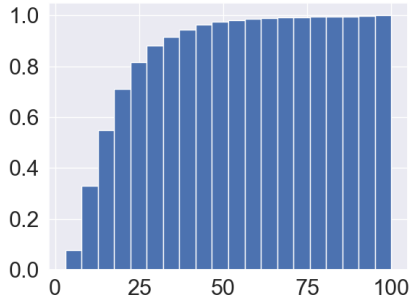


Fig. 2. Cumulative distribution of sentence length in the given labeled sentence set. The y axis represents the portion over all sentences.

(labeled with 1 or 0). For each binary classification task, given a set of labeled sentences consisting of n samples $S = \{s_1, \dots, s_n\}$ with labels $L = \{l_1, \dots, l_n\}$, where $l_i \in \{0, 1\}^2$, and a set of unlabeled sentences $S_u = \{s_1^u, \dots, s_m^u\}$, our goal is to learn the classifier $f(\hat{l}|s, \theta_i), i \in [1, 6]$. Our SMDA model contains several components: Supervised Learning (Section 4.1) for labeled sentences, Unsupervised Learning (Section 4.2) for unlabeled sentences, and Semi-Supervised Objective Function (Section 4.3) to combine labeled and unlabeled sentences.

4.1 Supervised Learning

Generating Balanced Labeled Training Set As shown in Fig. 1, the distribution is very unbalanced with respect to *General_support*, *Info_support* and *Emo_support*. In order to get more training sentences with these three types of support and make these three binary classification sub-tasks learn-able with a more balanced training set, we performed data augmentations over sentences with these three labels. Specifically, we paraphrased each sentence by 4 times via back translations and regarded that the augmented sentences have the same labels as original sentences. The comparison distributions are shown in Fig. 3

Supervised Learning for Labeled Sentences For each input labeled sentence s_i , we used XLNet [21] $g(\cdot)$ to encode it into hidden representation $h_i = g(s_i)$, and then passed them through a 2-layer MLP to predict the class distribution $\hat{l}_i = f(h_i)$. Since these sentences have specific labels, we optimize the cross entropy loss as supervised loss term:

$$L_S(s_i, l_i) = - \sum l_i \log f(g(s_i)) \quad (1)$$

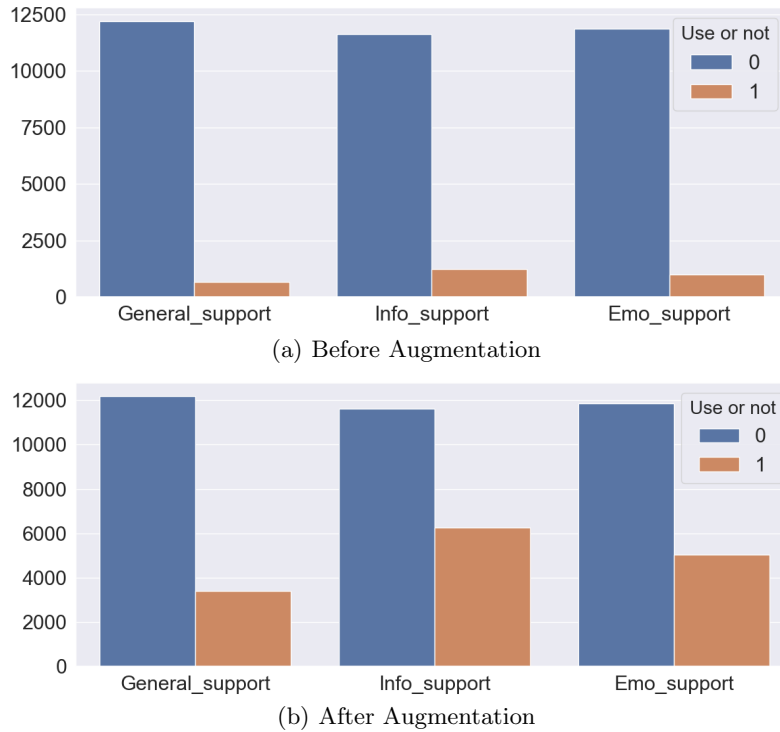


Fig. 3. Distributions before and after performing augmentations over labeled sentences belonging to *General_support*, *Info_support* and *Emo_support*. 0 means sentences don't use corresponding types of support, while 1 represents sentences use corresponding types of support. y axis is the number of sentences.

4.2 Unsupervised Learning

Paraphrasing Unlabeled Sentences We first performed back translations once for each unlabeled sentence $s_i^u \in S_u$ to generate the augmented sentence set $S_{u,a} = \{s_1^{u,a}, \dots, s_m^{u,a}\}$ in the same manner we described before.

Guessing Labels for Unlabeled Sentences For an unlabeled sentence s_i^u , we utilized $g(\cdot)$ and $f(\cdot)$ in Section 4.1 to predict the class distribution:

$$\hat{l}_i^u = f(g(s_i^u)) \quad (2)$$

To avoid the prediction being so close to uniform distribution, we generate low-entropy guessing labels \tilde{l}_i^u by a sharpening function [1]:

$$\tilde{l}_i^u = \frac{(\hat{l}_i^u)^{\frac{1}{\tau}}}{\|(\hat{l}_i^u)^{\frac{1}{\tau}}\|_1} \quad (3)$$

where $\|\cdot\|_1$ is l_1 -norm of the vector. When $T \rightarrow 0$, the guessed label becomes an one-hot vector.

Self-training for Original Sentences Inspired by self-training where model is also trained over unlabeled data with high-confidence predictions as their labels, in SMDA, with our guessed labels \tilde{l}_i^u with respect to original unlabeled sentence s_i^u , we added such pair (s_i^u, \tilde{l}_i^u) into training by minimize the KL Divergence between them:

$$L_s(s_i^u) = KL(f(g(s_i^u))\|\tilde{l}_i^u) \quad (4)$$

Entropy Minimization for Original Sentences One common assumption in many semi-supervised learning methods is that a classifier’s decision boundary should not pass through high-density regions of the marginal data distribution [5]. Thus for original unlabeled sentence s_i^u , we added another loss term to minimize the entropy of model’s output:

$$L_e(s_i^u) = - \sum f(g(s_i^u)) \log f(g(s_i^u)) \quad (5)$$

Consistency Regularization for Augmented Sentences With the assumption that the model should predict similar distributions with input sentences before and after augmentations, we minimized the KL Divergence between outputs with original sentence s_i^u as input and augmented sentence $s_i^{u,a}$ as input:

$$L_c(s_i^u) = KL(\hat{l}_i^u\|f(g(s_i^{u,a}))) \quad (6)$$

Combining all the loss terms for unlabeled sentences, we defined our unsupervised loss terms as:

$$L_U(s_i^u) = L_s(s_i^u) + L_e(s_i^u) + L_c(s_i^u) \quad (7)$$

4.3 Semi-Supervised Objective Function

We combined the supervised and unsupervised learning described above to form our overall semi-supervised objective function:

$$L = \mathbb{E}_{(s_i, l_i) \in (S, L)} L_S(s_i, l_i) + \gamma \mathbb{E}_{s_i^u \in S_u} L_U(s_i^u) \quad (8)$$

where γ is the balanced weight between supervised and unsupervised loss term.

5 Experiments

5.1 Model Setup

In SMDA ⁵, we only used single model for each task without jointly training and parameter sharing. That is, we trained six separate classifiers on these tasks.

⁵ The codes and data split will be released later.

Inspired by recent success in pre-trained language models, we utilized the pre-trained weights of XLNet and followed the same fine-tuning procedure as XLNet. We set the initial learning rate for XLNet encoder as $1e-5$ and other linear layers as $1e-3$. The batch size was selected in $\{32, 64, 128, 256\}$. The maximum number of epochs is set as 20. Hyper-parameters were selected using the performance on development set. The sharpen temperature T was selected in $\{0.3, 0.5, 0.8\}$ depending on different tasks. The balanced weight γ between supervised learning loss and unsupervised learning loss term started from a small number and grew through training process to 1.

5.2 Results

Our experimental results are shown in Table 3. We compared our proposed SMDA with BERT and XLNet in terms of accuracy(%) and Macro F1 score. BERT and XLNet achieved similar performance since they both obey the pre-training and fine-tuning manner. When combining with augmented and more balanced labeled data, massive unlabeled data, our SMDA achieved best performance across six binary-classification tasks. And we submitted the classification results on given unlabeled test set.

Table 3. Results on test set. Our baseline is our implementation of XLNet-cased-base.

Task	Emo_disc		Info_disc		Support		Gen_supp		Info_supp		Emo_supp	
	acc	F1	acc	F1	acc	F1	acc	F1	acc	F1	acc	F1
BERT _{BASE}	71.3	65.7	71.1	68.7	81.9	75.6	90.6	63.9	88.9	69.8	92.9	73.8
XLNet _{BASE}	72.4	67.9	72.2	69.3	83.4	77.3	92.7	65.0	87.9	70.3	93.4	73.8
SMDA	75.2	68.5	74.3	71.0	83.5	77.7	91.7	63.7	89.9	70.5	93.6	76.2

6 Conclusion

In this work, we focused on identifying disclosure and supportiveness in conversation responses based on a small labeled and large unlabeled training data via our proposed semi-supervised text classification system : Semi-Supervised Models via Data Augmentation (SMDA). SMDA utilized supervised learning over labeled data and conducted self-training, entropy minimization and consistency regularization over unlabeled data. Experimental results demonstrated that our system outperformed baseline models significantly.

References

- Berthelot, D., Carlini, N., Goodfellow, I.J., Papernot, N., Oliver, A., Raffel, C.: Mixmatch: A holistic approach to semi-supervised learning. CoRR [abs/1905.02249](https://arxiv.org/abs/1905.02249) (2019)

2. Chen, M., Tang, Q., Livescu, K., Gimpel, K.: Variational sequential labelers for semi-supervised learning. In: Proc. of EMNLP (2018)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019)
4. Ernala, S.K., Rizvi, A.F., Birnbaum, M.L., Kane, J.M., De Choudhury, M.: Linguistic markers indicating therapeutic outcomes of social media disclosures of schizophrenia. Proceedings of the ACM on Human-Computer Interaction **1**(CSCW), 43 (2017)
5. Grandvalet, Y., Bengio, Y.: Semi-supervised learning by entropy minimization. In: Proceedings of the 17th International Conference on Neural Information Processing Systems. pp. 529–536. NIPS’04, MIT Press, Cambridge, MA, USA (2004)
6. Gururangan, S., Dang, T., Card, D., Smith, N.A.: Variational pretraining for semi-supervised text classification. CoRR **abs/1906.02242** (2019)
7. Howard, J., Ruder, S.: Universal language model fine-tuning for text classification. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 328–339. Association for Computational Linguistics, Melbourne, Australia (Jul 2018)
8. Jaidka, K., Singh, I., Jiahui, L., Chhaya, N., Ungar, L.: A report of the CL-Aff OffMyChest Shared Task at Affective Content Workshop @ AAAI. In: Proceedings of the 3rd Workshop on Affective Content Analysis @ AAAI (AffCon2020). New York, New York (February 2020)
9. Kumar, A., Bhattamishra, S., Bhandari, M., Talukdar, P.: Submodular optimization-based diverse paraphrasing and its effectiveness in data augmentation. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 3609–3619. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). <https://doi.org/10.18653/v1/N19-1363>, <https://www.aclweb.org/anthology/N19-1363>
10. Lample, G., Conneau, A.: Cross-lingual language model pretraining. CoRR **abs/1901.07291** (2019)
11. Lee, D.H.: Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. ICML 2013 Workshop : Challenges in Representation Learning (WREPL) (07 2013)
12. Meng, Y., Shen, J., Zhang, C., Han, J.: Weakly-supervised neural text classification. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management. pp. 983–992. CIKM ’18, ACM, New York, NY, USA (2018). <https://doi.org/10.1145/3269206.3271737>, <http://doi.acm.org/10.1145/3269206.3271737>
13. Miyato, T., Dai, A.M., Goodfellow, I.: Adversarial training methods for semi-supervised text classification. In: International Conference on Learning Representations (2017)
14. Miyato, T., Maeda, S., Koyama, M., Ishii, S.: Virtual adversarial training: A regularization method for supervised and semi-supervised learning. IEEE Trans. Pattern Anal. Mach. Intell. **41**(8), 1979–1993 (2019)
15. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Proceedings of the 2018 Confer-

- ence of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 2227–2237. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018)
16. Warriner, A.B., Shore, D.I., Schmidt, L.A., Imbault, C.L., Kuperman, V.: Sliding into happiness: A new tool for measuring affective responses to words. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale* **71**(1), 71 (2017). <https://doi.org/10.1037/cep0000112>, <https://app.dimensions.ai/details/publication/pub.1084126691andhttp://europepmc.org/articles/pmc5334777?pdf=render>
 17. Wei, J.W., Zou, K.: EDA: easy data augmentation techniques for boosting performance on text classification tasks. *CoRR* **abs/1901.11196** (2019)
 18. Xie, Q., Dai, Z., Hovy, E., Luong, M.T., Le, Q.V.: Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848* (2019)
 19. Yang, D., Kraut, R.E., Smith, T., Mayfield, E., Jurafsky, D.: Seekers, providers, welcomers, and storytellers: Modeling social roles in online health communities. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. p. 344. ACM (2019)
 20. Yang, D., Yao, Z., Seering, J., Kraut, R.: The channel matters: Self-disclosure, reciprocity and social support in online cancer support groups. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. p. 31. ACM (2019)
 21. Yang, Z., Dai, Z., Yang, Y., Carbonell, J.G., Salakhutdinov, R., Le, Q.V.: Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR* **abs/1906.08237** (2019)
 22. Yang, Z., Hu, Z., Salakhutdinov, R., Berg-Kirkpatrick, T.: Improved variational autoencoders for text modeling using dilated convolutions. *CoRR* **abs/1702.08139** (2017)
 23. Zhang, P.: The affective response model: A theoretical framework of affective concepts and their relationships in the ict context. *Management Information Systems Quarterly (MISQ)* **37**, 247–274 (03 2013). <https://doi.org/10.25300/MISQ/2013/37.1.11>