Label Smoothing for Emotion Detection (Student Abstract)

George Maratos, Tiberiu Sosea, Cornelia Caragea

Computer Science
University of Illinois at Chicago
{gmarat2, tsosea2, cornelia}@uic.edu

Abstract

Automatically detecting emotions from text has countless applications, ranging from large scale opinion mining to social robots in healthcare and education. However, emotions are subjective in nature and are often expressed in ambiguous ways. At the same time, detecting emotions can also require implicit reasoning, which may not be available as surface-level, lexical information. In this work, we conjecture that the overconfidence of pre-trained language models such as BERT is a critical problem in emotion detection and show that alleviating this problem can considerably improve the generalization performance. We carry out comprehensive experiments on four emotion detection benchmark datasets and show that calibrating our model predictions leads to an average improvement of 1.35% in weighted F1 score.

Introduction

Improving existing approaches for automatic emotion detection is part of a larger goal to design machines that can identify affective patterns in our behavior (Picard 2000) and to better understand human emotion. However, as prior work has shown, people can express emotions in an implicit manner. In this case, lexical cues indicative of the conveyed emotion are absent from the text, hence models of emotion detection need a deeper understanding in order to correctly identify the expressed emotion. Moreover, learning useful representations for implicit text requires sufficient amounts of labeled data, which may not be readily available.

In this work, we focus on various challenges posed by implicit emotions in low-resource settings and perform an extensive analysis to contrast implicit expressions of emotions against explicit expressions of emotions. Interestingly, we find that the model performance degrades significantly when dealing with implicit text, especially when labeled data is scarce. Next, we argue that the overconfidence of pre-trained language models is a critical issue in this setup. Emotions may be ambiguous and the task is subjective, hence we argue that the model should yield *softer* predictions to account for the nature of the task itself. Moreover, we postulate that overconfidence may also lead to a blind reliance on lexical cues. Consider the following example: *I love that I have to pay triple now*. Note that in this example, the *annoyance* of the writer is expressed implicitly. However, the presence of the word *love* may mislead the model, especially if it has been trained to confidently associate the *love* lexical cue with the *love* emotion.

To overcome these challenges, we propose to leverage Label Smoothing Regularization (LSR) (Szegedy et al. 2016). LSR converts the sharp (one-hot) target label into a mixture of the one hot and the uniform distribution, and trains the model on the resulting soft distribution. We show that LSR improves the performance of BERT (Devlin et al. 2019) for both explicit and implicit text. Critically, we observed that improvements on implicit text is greater than that on explicit text, which shows that training with soft labels yields better language representations for emotion detection in low resource settings.

Methodology

Datasets and Pre-processing We consider four emotion detection datasets with various emotion taxonomies: EmoInt (Mohammad and Turney 2013) (4 emotions), ISEAR (Scherer and Wallbott 1994) (7 emotions), GoEmotions (Demszky et al. 2020) (27 emotions), and Empathetic Dialogues (Rashkin et al. 2019) (32 emotions). We also drop multi-label examples and the neutral class for the GoEmotions dataset.

Explicit and Implicit Emotions in Text In this work, an *explicit* example is any text that contains lexical cues indicative of the conveyed emotion. For example, *I feel guilty* is an explicit example since the *guilt* emotion is expressed at a lexical level using the word *guilty*. In contrast, implicit text is any text where lexical cues are absent. To identify implicit or explicit text, we use the EmoLex (Mohammad and Turney 2013) emotion-word association lexicon, which associates words with one of the eight Plutchik-8 emotions. Concretely, given a piece of text t annotated with an emotion e, we consider t to be explicit if there is any word w from t that is annotated with emotion e in EmoLex, and implicit otherwise.

Experimental Procedure We measure the performance of our LSR-trained models along both weighted F1 and expected calibration error. The calibration error measures how accurately a network's probability estimates reflect the true correctness likelihood. Accurately calibrated predictions are particularly important in applications of emotion

		ALL DATA					EXPLICIT					IMPLICIT				
		10	25	50	100	AVG	10	25	50	100	AVG	10	25	50	100	AVG
Emp. Dia.	BERT	34.5	43.4	46.4	49.1	43.5	39.5	48.9	52.0	53.5	48.5	32.5	41.0	43.9	47.1	41.1
	BERT + LSR	36.1†	44.4†	47.2†	50.1†	44.5	41.3	49.7	52.3	54.6	49.5	34.0	42.2	44.9	48.1	42.3
GoEmotions	BERT	40.2	51.3	54.9	56.6	50.8	42.6	54.4	58.2	59.9	53.8	39.9	50.7	54.2	55.9	50.2
	BERT + LSR	42.2†	52.5†	55.4†	57.6†	52.0	44.1	55.2	58.3	60.1	54.4	42.1	52.0	54.9	57.1	51.5
EmoInt	BERT	32.5	44.1	62.2	76.2	53.8	31.5	41.2	61.2	71.3	51.3	20.4	23.5	27.8	34.0	26.4
	BERT + LSR	34.6†	47.1†	63.8	78.1†	55.9	35.5	44.8	62.3	72.0	53.7	21.2	25.2	28.1	35.3	27.5
ISEAR	BERT	37.0	53.7	59.5	63.2	53.4	33.5	58.8	64.3	67.7	56.1	37.8	53.2	58.9	62.7	53.2
	BERT + LSR	39.1†	54.3	60.4	64.0†	54.5	36.5	57.9	66.3	68.9	57.4	39.8	54.0	59.7	63.4	54.2

Table 1: Weighted F1 scores of our models. The AVG column is the average across all subsets of the data. The † indicates that the difference in performance between BERT and BERT+LSR for that subset of the data is statistically significant ($p<0.05$).

		ALL DATA					EXPLICIT					IMPLICIT				
		10	25	50	100	AVG	10	25	50	100	AVG	10	25	50	100	AVG
Emp. Dia.	BERT	23.3	33.0	28.9	23.1	27.1	22.2	31.5	26.9	21.4	25.5	26.0	36.6	33.6	26.9	30.8
	BERT + LSR	10.5	7.7	11.3	9.2	9.7	11.1	8.8	11.2	10.6	10.4	10.1	6.7	12.9	7.4	9.3
GoEmotions	BERT	23.4	31.6	30.5	27.1	28.15	20.8	28.0	26.9	23.6	24.8	29.0	39.9	39.3	35.8	36.0
	BERT + LSR	9.7	8.1	11.3	14.9	11.0	11.0	11.0	12.9	16.7	12.9	9.1	5.2	10.5	12.1	9.2
EmoInt	BERT	35.8	37.7	28.3	18.4	30.1	34.8	35.8	24.3	14.5	27.4	42.0	49.3	52.9	42.6	46.7
	BERT + LSR	13.2	14.9	7.9	13.2	12.3	12.5	13.9	8.1	14.3	12.2	20.2	26.2	24.2	15.3	21.5
ISEAR	BERT	36.6	28.7	30.0	27.4	30.7	33.6	23.3	24.1	21.4	25.6	43.0	41.7	43.4	41.8	42.5
	BERT + LSR	8.1	11.2	8.8	10.7	9.7	7.2	10.6	9.3	13.3	10.1	13.2	17.3	13.4	10.8	13.7

Table 2: The expected calibration error on all emotion detection tasks.

detection, such as empathetic conversational chatbots where confidently confusing *joy* for *sadness* might lead to harmful responses. - this could be commented out For each dataset, we train our models on random subsets of the data to simulate a low resource setting. We construct our subsets by sampling the data evenly among the classes, and we explore 4 low resource settings with $n = \{10, 25, 50, 100\}$, where n is the number of examples per class. After training, we evaluate all models on the test set provided by each dataset, which we divide into an implicit group and an explicit group using the procedure shown in the previous paragraph. We report the performance of our models on the implicit test set, explicit test set, as well as the original test set.

Main Results and Conclusion

We show the results of our experiments in terms of weighted F1 score in Table 1 and make two observations: **1) LSR can significantly improve performance in low resource settings** with an average improvement of 1.35% over the vanilla BERT model. Notably, on EmoInt, LSR pushes the performance by as much as 2.1% in F1, indicating the effectiveness of smooth labels. **2) LSR leads to larger improvements on implicit examples over explicit.** For example, using 100 examples per class on GoEmotions, the improvement on the explicit set is only 0.2% compared to 1.2% on the implicit set. Note that detecting implicit emotions is considerably more challenging than explicit emotions, which shows that soft labels lead to more powerful representations for emotion detection.

We also show in Table 2 the performance of our models in terms of expected calibration error (ECE), where we observe similar trends. Here, the gap in ECE between BERT and BERT+LSR models on the implicit set is much larger than on the explicit set. This suggests that **using LSR during training will lead to substantially better calibrated models on implicit text**.

In conclusion, we show that LSR is a simple way to improve the performance and calibration of BERT in the low resource setting for emotion detection.

Acknowledgements

This research is supported in part by NSF Convergence Accelerator award #2137846, NSF IIS award #2107487, and NSF BigData award #1912887.

References

Demszky, D.; Movshovitz-Attias, D.; Ko, J.; Cowen, A.; Nemade, G.; and Ravi, S. 2020. GoEmotions: A Dataset of Fine-Grained Emotions. In *ACL*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*.

Mohammad, S. M.; and Turney, P. D. 2013. Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, 436–465.

Picard, R. W. 2000. *Affective computing*. MIT press.

Rashkin, H.; Smith, E. M.; Li, M.; and Boureau, Y.-L. 2019. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. In *ACL*.

Scherer, K. R.; and Wallbott, H. G. 1994. Evidence for universality and cultural variation of differential emotion response patterning. *JPSP*.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *CVPR*, 2818–2826.