Detecting Optimism in Tweets using Knowledge Distillation and Linguistic Analysis of Optimism

Ștefan Cobeli*, Bogdan Iordache†, Shweta Yadav*, Cornelia Caragea*,
Liviu Dinu†, Dragoș Iliescu†

*University of Illinois at Chicago, †University of Bucharest

Abstract

Finding the polarity of feelings in texts is a far-reaching task. Whilst the field of natural language processing has established sentiment analysis as an alluring problem, many feelings are left uncharted. In this study, we analyze the *optimism* and *pessimism* concepts from Twitter posts to effectively understand the broader dimension of psychological phenomenon. Towards this, we carried a systematic study by first exploring the linguistic peculiarities of optimism and pessimism in user-generated content. Later, we devised a multi-task knowledge distillation framework to simultaneously learn the target task of optimism detection with the help of the auxiliary task of sentiment analysis and hate speech detection. We evaluated the performance of our proposed approach on the benchmark Optimism/Pessimism Twitter dataset. Our extensive experiments show the superiority of our approach in correctly differentiating between optimistic and pessimistic users. Our human and automatic evaluation shows that sentiment analysis and hate speech detection are beneficial for optimism/pessimism detection.

Method

Knowledge Distillation Setting Given a dataset $\mathcal{D}$, a student model $\mathcal{S}$ learns only to mirror the logits generated by a teacher $\mathcal{T}$, minimizing the cross entropy loss between the outputs of the student model and the outputs of the teacher model:

$$\mathcal{L}_{\text{KD}} = \sum_{(x,y)\in\mathcal{D}} \ell(f(x;\theta_{\mathcal{S}}), f(x;\theta_{\mathcal{T}})).$$

Loss with respect to a Single Teacher: We augmented the patient knowledge loss to use *Teacher Annealing*:

$$\mathcal{L}_{\mathcal{T}} = (1-\alpha) * \mathcal{L}_{01} + \alpha * (\mathcal{L}_{\text{KD}} + \beta * \mathcal{L}_{\text{PKD}}).,$$

α linearly decreases from 1 towards 0.

Loss with respect to Multiple Teachers: Given the tasks $\{\mathcal{T}_1, \ldots, \mathcal{T}_t\}$, we define the multi-task loss as:

$$\mathcal{L}_{\text{MTKD}} = \sum_{i\in 1,t} \mathcal{L}_{\mathcal{T}_i}.$$

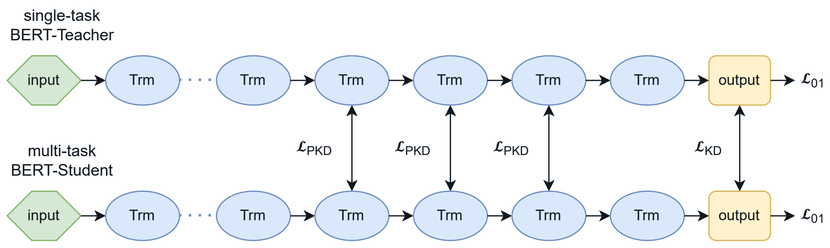


Fig. 1: Proposed distillation architecture from a teacher model (top) to a student model (bottom).

Datasets

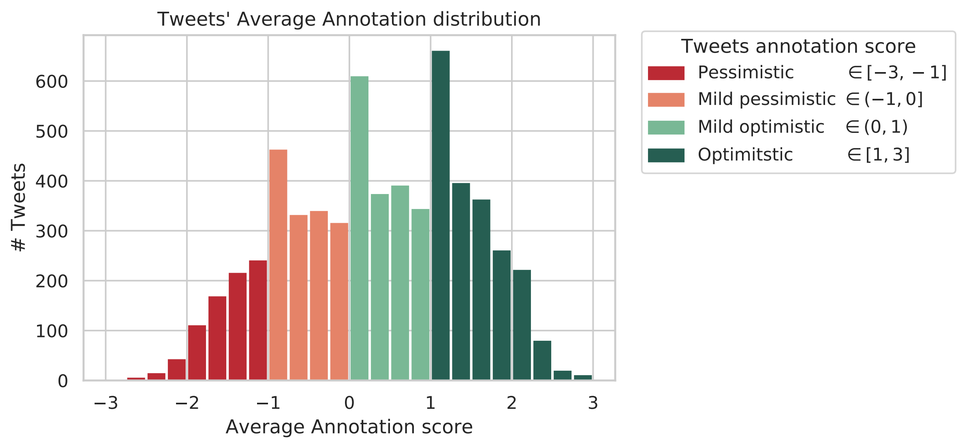


Fig. 2: Each tweet was rated by 5 different annotators with a score $\in\{-3,\ldots,3\}$, according to the tweet's optimism level. The average annotation score was obtained by averaging the 5 individual scores [RWM16].

Hate Speech Dataset (Hate) is a dataset developed for the analysis of offensive conducts in Twitter posts [Fou+18]. There are $80,000$ tweets each one labeled as either *normal* (59%), *spam* (22.5%), *abusive* (11%) or *hateful* (7.5%).

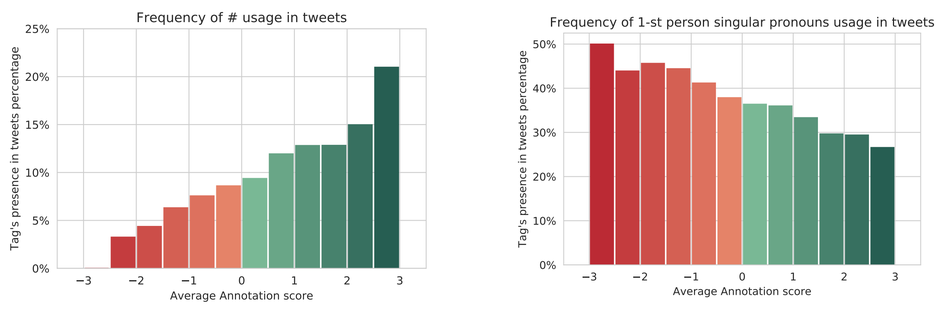
Sentiment Polarity Dataset (Sent) is a twitter sentiment dataset (Sent) proposed at the SemEval competition in 2017 [RFN17]. The Sent dataset is composed of $50,333$ tweets annotated with one of the three labels: *negative* (15.57%), *neutral* (44.89%) or *positive* (39.54%).

Results

Model	Test Acc.	Dev Acc.
BERTweet	84.84	84.58
MTKD OPT + Hate	86.52	85.30
MTKD OPT + Sent	86.23	85.44
MTKD OPT + Hate + Sent	**86.60**	85.14
MTKD no KD	82.11	81.82
MTKD vanilla-BERT	85.64	84.71
MTKD downsampled	86.19	85.23
XLNet Base	84.25	–
BERT Base with SLA [Als+21]	85.69	–

Tab. 1: Best models' performances on optimism prediction. We can see that all the components of the best model are relevant. The combination between BERTweet, PKD and both intermediate tasks provides the highest accuracy on the test set.

Analysis



(a) Hashtags (b) First Person Pronouns

Fig. 3: Usage frequency of hashtags (a) and of first person singular pronouns (b) in tweets as the optimism polarity increases from -3 to $+3$.

	Tweet	Average Annotation	Prediction Confidence
Original tweet	i don't know how they can be so perfect	-0.2	55.03%
Pessimistic correction	i don't know how they can be so perfect **liars!**	-1.2	77.01%
Optimistic correction	**flawless!** i don't know how they can be so perfect.	1.0	85.03%

Tab. 2: Tweet modified to be more optimistic and pessimistic. Whilst the original tweet was misclassified by our best model, after limited clarifying corrections the model predicts accurately the pessimism and respectively the optimism of the tweet.

Conclusions and Future Work

- We have presented a multi task learning methodology for optimism detection and highlighted a relationship that can be explored between optimism, sentiment and mental health;
- We have explored linguistic characteristics expressed in optimism and pessimism;
- The relationship between optimism-pessimism and mental health constructs can be further explored.

Selected References

[Als+21] Ali Alshahrani et al. "Optimism/Pessimism Prediction of Twitter Messages and Users Using BERT with Soft Label Assignment". In: *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2021, pp. 1–8.

[Fou+18] Antigoni Maria Founta et al. "Large scale crowdsourcing and characterization of twitter abusive behavior". In: *Twelfth International AAAI Conference on Web and Social Media*. 2018.

[RFN17] Sara Rosenthal, Noura Farra, and Preslav Nakov. "SemEval-2017 task 4: Sentiment analysis in Twitter". In: *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*. 2017, pp. 502–518.

[RWM16] Xianzhi Ruan, Steven Wilson, and Rada Mihalcea. "Finding optimists and pessimists on twitter". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 2016, pp. 320–325.