

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, Illinois, US

†University of Bucharest, Bucharest, Romania

*{scobel2, shwetay, cornelia}@uic.edu

†{iordache.bogdan1998@gmail.com, liviu.p.dinu@gmail.com, dragoș.iliescu@fpse.unibuc.ro}

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.

1. Introduction

The optimism-pessimism continuum is a stable personality trait, represented by the dispositional tendency to hold generalized and positive expectancies even when confronted with adversity and stress (Scheier et al., 2001). The optimism-pessimism continuum has been intensively studied because of its links to a number of important outcomes, both in the realm of performance (e.g., job performance; (Kumar et al., 2017)), and satisfaction (e.g., life satisfaction; (Extremera et al., 2009)). Among all the individual differences, optimism-pessimism is one of the strongest connected to happiness (Alarcon et al., 2013), stress management (Carver et al., 2010) and ultimately mental (Scheier and Carver, 1992) and, perhaps unexpectedly, physical wellbeing (Rasmussen et al., 2009). The mechanisms of this connection are not yet all clear, but we know now that optimism has a versatile role in different context: it acts as a strong protective factor against adversity (Gallagher et al., 2020), it acts as a distal predictor and encourages effective coping against and recovery from anxiety, depression, traumatic events and suicidal inclinations (Achat et al., 2000; Prati and Pietrantonio, 2009), and it has a strong cross-over effect to close relatives and acquaintances (Scheier et al., 2001; Peterson and Bossio, 2001).

The fundamental mindset, i.e., the generalized positive or negative expectancy, or "frame" (McKenzie and Nelson, 2003), held by a person who is characterized by optimism or pessimism influences in a profound manner the ways in which that person observes, processes, understands and reacts to the environment. Knowledge of this mindset therefore permits both immediate and long term predictions of human behavior. Assessing dispositional optimism-pessimism on the fly is there-

fore of significant practical importance. The methods developed so far for the assessment of dispositional optimism-pessimism are however tributary to the classical psychometric view (i.e., are mostly self-report measures) and are therefore difficult to deploy and are slow in any computer-mediated interaction. Faster and automatic assessment methods based on a computational approach are needed to cover this practical gap.

From a computational perspective, optimism and pessimism were studied first by Ruan et al. (2016) using only standard classifiers such as Naïve Bayes, without a focus on semantics. Later, Caragea et al. (2018) provided a first analysis of optimism/pessimism using deep learning models and showed the role of analyzing the sentiments in optimism/pessimism detection.

Besides the sentiment, abusive language (hate speech) and attitudes are also highly associated with the emotional and psychological state of the speaker (Patrick, 1901), that is contemplated in the affective attributes of their language (Mabry, 1974). In this paper, we aim to model both phenomena jointly in a multi-task learning paradigm to accurately predict optimistic/pessimistic users. Specifically, we utilized the concept of knowledge distillation (KD) (Hinton et al., 2015; Clark et al., 2019), where we aim to effectively transfer the knowledge from multiple teacher networks to a student network. Our proposed knowledge distillation framework is trained with multiple distillation techniques (vanilla KD, patient KD, and teacher annealing), which helps the student model imitate the teacher's behavior effectively.

This paper presents the following contribution:

- We conducted an in-depth analysis on the linguistic characteristics of pessimism and optimism.

- We proposed a novel multi-task knowledge distillation framework to improve optimism/pessimism prediction by transferring knowledge from other complementary tasks of sentiment analysis and hate speech detection.
- Our experiment demonstrates the effectiveness of the proposed approach in significantly improving over transfer learning and state-of-the-art methods on the benchmark Optimism/Pessimism Twitter dataset.
- Our human analysis confirms our hypothesis of the importance of sentiment analysis and hate speech detection in detecting optimistic and pessimistic users.

2. Related Work

Social media platforms such as Twitter or Facebook have become environments where people come together to share their thoughts, opinions, ideas - and in a more covert manner their feelings, attitudes, sentiments and emotions. Online behaviors are loaded and entangled with feelings.

While sentiments and emotions have been studied extensively in social media (Go et al., 2009; Ortigosa et al., 2014; Waterloo et al., 2018), surprisingly, the traits or dispositions that underlie emotions have received very little attention. Positive-negative emotion is the most basic manner in which emotions can be described - and this split is largely tributary on a personality trait level to the optimism-pessimism divide: an important difference in many applications. Personality traits are the basis of, and significantly influence, most online human behavior. For example, when an individual writes an opinion or a product review, these are affected not only by the user experience and product interaction, but also, by that individual’s personality and attitudes. Dispositionally pessimistic users may use extremely positive ratings (e.g., excellent) less frequently, which will affect the overall rating of a product. Pessimism is also highly correlated with negative affect, feelings of vulnerability, depressed moods and even clinical depression (Yan and Tan, 2014; Qiu et al., 2011). Pessimism is therefore both a vulnerability factor and an early warning flag for mental health issues, and early detection of pessimism may help provide the necessary support to individuals who are at risk to later develop subclinical patterns or clinical symptoms. The early and fast detection of optimism-pessimism is therefore an important objective, with a high practical and social value.

Recent advances in natural language processing show that the large pre-trained language models such as BERT (Devlin et al., 2018) or XLNet (Yang et al., 2019) yield substantial improvements in performance in many downstream tasks. Moreover, further fine-tuning of these models on related data-rich intermediate tasks generates additional improvement in perfor-

mance on many target tasks (Han and Eisenstein, 2019; Pruksachatkun et al., 2020). Optimism-pessimism detection was approached from classical machine learning perspective by Ruan et al. (2016) and from a deep learning perspective by Caragea et al. (2018). By using XLNet the pessimism-optimism detection performance was further improved by Alshahrani et al. (2020). As well, further improvement was obtained by using BERT along with soft label assignments, in (Alshahrani et al., 2021). We extend ongoing expertise on optimism-pessimism detection by using information rich attention based models pre-trained on Twitter posts. As well, we show that multi-task knowledge distillation reveals a significant improvement for the optimism detection task.

Optimism detection is a problem in which annotated data are lacking. Therefore, the usage of pre-trained models is essential from a deep learning point of view. The advantage of pre-trained models is that they have already learned linguistic structures by being trained to solve multiple tasks (Radford et al., 2019). Recently, it has been shown that by using sufficiently large language models, any downstream task can become a few-shot learning problem (Brown et al., 2020). There are multiple alternatives to the transfer of knowledge between tasks. Usually, unsupervised objectives are employed to this end, such as the prediction of the next word in a sentence or the placement of a number of shuffled words in a meaningful order (Raffel et al., 2020). In this paper, we address the problem of lack of annotated data through intermediate-tasks knowledge transfer using pre-trained language models for optimism and pessimism detection.

3. Method

Our proposed multi-task knowledge distillation framework aims to transfer knowledge from models trained on other complementary tasks, such that the resulting model has the ability to discriminate between optimism and pessimism taking advantage on the secondary tasks’ additional knowledge. A common pitfall of multitask learning is that often the joint model has a lower performance on the individual tasks than single task trained models (Martínez Alonso and Plank, 2017). To address this problem we follow a multi task knowledge distillation methodology. Using multiple specialized teachers and distilling the knowledge to student model has been shown to proven to counter this issue (Clark et al., 2019).

We begin by first describing our approach to learn optimism with the help of a single teacher model and later expanding it to a multiple teacher.

3.1. Learning Optimism with the help of a teacher

A novel multi-task learning approach implies the concept of knowledge distillation (KD) (Hinton et al., 2015; Clark et al., 2019). In order to build multi-task

knowledge distillation framework, we first analyze how single task knowledge distillation can be applied to our problem. The concept of knowledge distillation seeks to smoothen the learning phase. First, a teacher model is trained on the target dataset. Second, a student model is trained to mimic the outputs of the trained teacher. Thus, the student model benefits from a training signal that is richer than a simple one-hot-encoded target.

Towards this, we investigated three distillation techniques: (i) *vanilla knowledge distillation* (KD) (Hinton et al., 2015), (ii) *patient knowledge distillation* (PKD) (Sun et al., 2019) and (iii) multi-task knowledge distillation (MTKD) adapted from (Clark et al., 2019).

In a vanilla KD, for 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 $\mathcal{L}_{\text{KD}} = \sum_{(x,y) \in \mathcal{D}} \ell(f(x; \theta_{\mathcal{S}}), f(x; \theta_{\mathcal{T}}))$. This approach can be further augmented by training the student to mimic a linear interpolation between teacher’s logits and the one-hot-encoded ground truth.

In the context of PKD, besides learning the teachers’ outputs, the student also learns a number of the teachers’ intermediate layers. Even though PKD was introduced as a model compression technique, we have successfully adapted it for teaching a student of the same size as the teacher. In our context we found that the student performs best when it learns to mirror the outputs of the last three layers of the teacher model.

In order to standardize both KD and PKD learning methods, for each teacher \mathcal{T} we developed an unitary loss function based on the loss between the one hot encoded labels and student’s output (\mathcal{L}_{01}), the loss between the teacher’s output and the student’s output (\mathcal{L}_{KD}) and the loss between the teacher’s and the student’s intermediate layers that we are mirroring (\mathcal{L}_{PKD}):

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

The hyperparameter β represents the importance of the PKD loss, whilst α controls the general impact of the distillation procedure. When we use teacher annealing (TA), we are linearly decreasing the value of the hyperparameter $\alpha \in (0, 1]$ towards 0. Thus, in the beginning of the training phase the student has access mostly to the distilled knowledge $\mathcal{L}_{\text{KD}} + \beta * \mathcal{L}_{\text{PKD}}$. As the student performance increases, the loss approximates the plain binary cross entropy loss \mathcal{L}_{01} based solely on the ground truth.

3.2. Learning Optimism with the help of multiple teacher

We proposed a multi-task knowledge distillation (MTKD) method adapted from (Clark et al., 2019). The MTKD methodology implies training single-task teacher models $\{\mathcal{T}_1, \dots, \mathcal{T}_t\}$, each on a different dataset $\{\mathcal{D}_1, \dots, \mathcal{D}_t\}$. A multi-task student model learns to mirror each teacher \mathcal{T}_i by minimizing the sum

of the knowledge distillation losses with respect to each teacher $\mathcal{L}_{\mathcal{T}_i}$. We thus adapt the distillation objective presented above for multiple tasks. We define the new MTKD student loss as the sum of the losses computed for each task $\mathcal{L}_{\text{MTKD}} = \sum_{i \in \overline{1,t}} \mathcal{L}_{\mathcal{T}_i}$.

Compared to the vanilla KD setting, the student has access to multiple training sources and it was shown to improve its performance on all tasks with respect to its teachers.

4. Datasets

In this section we first discuss the target and intermediate-tasks and corresponding datasets used in our transfer learning approach, and then refer to additional implementation details.

Target Task The primary dataset we consider in our study was Optimism/Pessimism Twitter (OPT) dataset, introduced by Ruan et al. (2016). The OPT dataset was collected from Twitter and consists of 7,475 tweets. Each tweet was rated by five individual annotators with an integer score between -3 (very pessimistic) and $+3$ (very optimistic). The golden standard rating for each tweet was considered to be the average of the five received scores. In previous studies, two settings for the definition of *optimism* and *pessimism* were considered in this context. The first setting considers the tweets with a golden standard smaller or equal to 0 as pessimistic and the tweets with positive golden standard as optimistic (i.e., the cut-off is place on the theoretical mean of 0). The second setting considers only the tweets with ratings smaller or equal to -1 (pessimistic) and with ratings greater or equal to 1 (optimistic), while ignoring the tweets between -1 and 1 (i.e., considers only the more pronounced rating to call the class of a tweet and ignores those that are clustered more towards the mean).

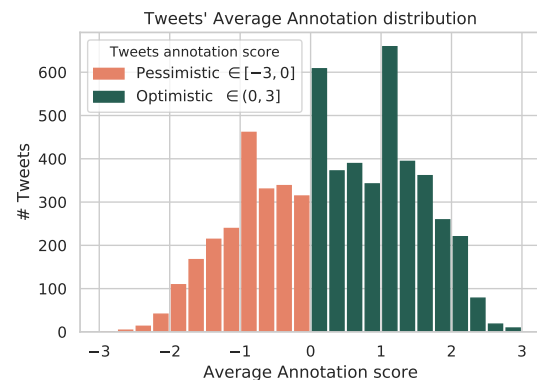


Figure 1: Distribution of optimism review scores in OPT dataset. Optimistic tweets (pictured in green) add up to 62.6% of the dataset. In this case, we considered as *pessimistic* the tweets annotated with a score of 0 or lower and as *optimistic* the tweets with a score larger than 0.

In our analysis we refer to the first setting as the 0

setting and to the latter as the 1/ - 1 setting. We are mainly investigating the best models for the optimism prediction in the 0 setting and evaluate them also on the 1/ - 1 setting. In Figure 1, we can visualize the distribution of the golden standard in the OPT dataset. We thus have 62.60% (4,679/7,475) optimistic tweets and 37.40% (2,796/7,475) pessimistic tweets in the 0 setting. Whilst, in the 1/ - 1 setting we have 65.17% (2,507/3,847) optimistic tweets and 34.83% (1,340/3,847).

Intermediate Tasks One of the behaviors that was intensely studied in social media datasets is the presence of toxicity in texts. We posit that *hate speech* may also have an impact in determining if a tweet is either optimistic or pessimistic. We therefore also used the dataset (*Hate*) introduced by (Founta et al., 2018). This dataset was labeled using an iterative procedure, in multiple rounds. There are 80,000 tweets each one labeled as either *normal* (59%), *spam* (22.5%), *abusive* (11%) or *hateful* (7.5%).

Finally, another association that was proposed by (Caragea et al., 2018) is between optimism and sentiment polarity. We used a twitter sentiment dataset (*Sent*) proposed at the SemEval competition in 2017 (Rosenthal et al., 2017). 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%).

All the datasets discussed above are based on tweets. We believe that training on a larger number of tweets may give the models a better understanding of the tweets’ linguistic structure.

4.1. Dataset Analysis: Linguistic Aspects of Optimism

We start our analysis by scrutinizing the usage of various parts of speech in pessimistic and optimistic tweets. In the context of social media texts we also take into account some structures that may not be canonically considered as parts of speech, such as hashtags or emoticons.

We started the lexical analysis by employing a Twitter-aware tokenizer and part-of-speech tagger (Owoputi et al., 2013) on the OPT dataset. This parser detects besides the traditional parts of speech (such as nouns or verbs) also social media platform specific elements (such as hashtags or emoticons). We will further refer to the features detected by this parser simply as *tags*.

By considering all the tweets annotated with a label greater than 0 to be optimistic and the others pessimistic, as proposed in the 0 setting definition, we are making a broad generalization. Nevertheless, we can still find linguistic features to discriminate between these two broad categories. For example, in the context of a pessimistic tweet there are more adverbs used than in an optimistic one, as can be seen in Figure 2.

Using 0 as threshold for considering tweets as either pessimistic or optimistic is a convention that ignores

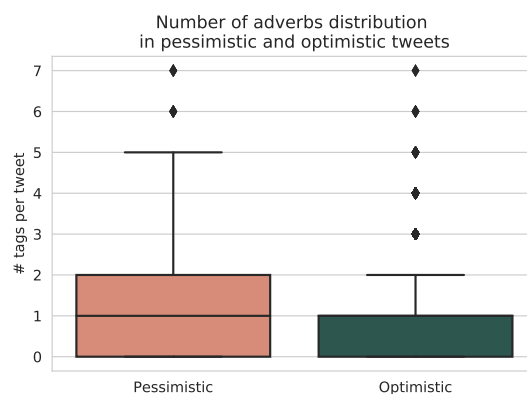


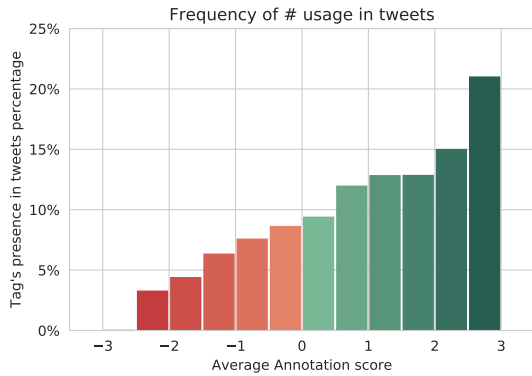
Figure 2: Distribution of the number of adverbs used in pessimistic and optimistic tweets. We can see that pessimistic tweets tend to have more adverbs than optimistic ones.

the fact that some texts or tweets have no optimism polarity; they may be formal or neutral. Thus, it is natural to further study how linguistic features behave throughout the entire pessimism-optimism continuous spectrum.

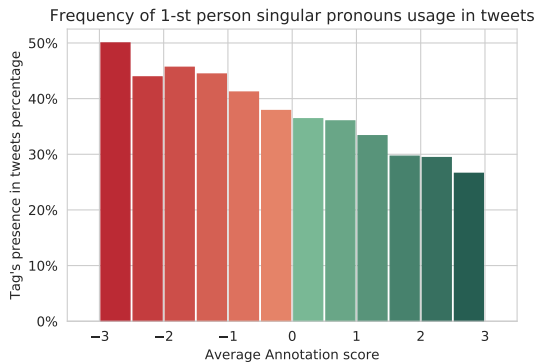
We studied the frequencies of the parts of speech (tags) usage as the optimism annotation label increases from -3 to 3. We noticed that some tags are uniformly common in any range of optimism polarity. For example, nouns appear in between 80% and 90% percent of the tweets independent of their optimism range. On the other hand, the probability of encountering hashtags in a tweet increases proportionally with its optimism level, as can be seen in Figure 3 (a). Other features that we noticed may be correlated with optimism and pessimism are the presence of emoticons, punctuation or user mentions.

One peculiarity that is especially intriguing is the usage of first person singular pronouns. As shown in Figure 3 (b), first person pronouns tend to be used less frequently as the tweets get more optimistic. This is a result that correlates with common intuition that pessimism might be linked with depression; this phenomenon was pointed out by (Zimmermann et al., 2017), who argued that self-focused attention, indicated by the use of first person singular pronouns, can predict depressive symptoms.

Based on all these observations and in order to inquire how well they might generalize we built a classifier based on the occurrences of these features. We represented each tweet as a vector of occurrences of length 26, corresponding to the tagset features expressed by the tweets parser (Owoputi et al., 2013). We trained an XGBoost regressor (Chen and Guestrin, 2016) on the vectorized OPT training set to predict the average annotation of each tweet’s optimism polarity. In order to test the prediction performance we discretized the prediction on the same 0 threshold as in the 0 setting optimism formulation. By doing so, we obtained a clas-



(a)



(b)

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

sifier with 6% better accuracy than a classifier which would predict always the most frequent class. This result reveals that linguistic features has the potential to be a relevant predictor in optimism detection.

The classifier based on these linguistic tags revealed that the most important features were the first person singular pronouns, the hashtags, the emoticons and the punctuation. On the opposite side, nouns since are the most common parts of speech, have a negligible discrimination importance. Nouns are present almost on the same contexts in both pessimism and optimism.

In this section we have seen that there is variation on the pessimism-optimism spectrum that may be explained by linguistic features. In the following sections we will investigate the potential of attention based models on optimism prediction.

4.2. Experimental Setup

In our deep learning experiments, we split the OPT dataset in three disjoint subsets (training, validation and test) of sizes 80%, 10% and 10% respectively, of the original. We report the results as the mean accuracy of optimism/pessimism prediction over five independent runs. The pre-trained language models that we utilized are BERT (Devlin et al., 2018) and BERTweet (Nguyen et al., 2020). For each of these models we

Model	Test Acc.	Dev. Acc.
BERT Base	83.90	82.62
BERTweet	84.84	84.58
KD	85.21	83.94
PKD	85.21	84.79
PKD and T.A.	85.45	84.85
GRUStack baseline (Caragea et al., 2018)	80.19	–

Table 1: Accuracy of BERT models combined with distillation techniques on the OPT dataset compared with the best non-BERT baseline.

used only the embedding of the [CLS] token, as suggested for the GLUE tasks.

We found that the optimal values for the hyperparameters were α with initial value 1 and β with value 100.

Our implementation was based on PyTorch (Paszke et al., 2019) and the main model architectures were those provided by the Hugging Face library (Wolf et al., 2019).

5. Results

To analyze the effectiveness of our approach, we compared our model with multiple baselines (BERT Base, BERTweet and GRUStack (Caragea et al., 2018)) as presented in Table 1. The results show that solely fine-tuning BERT on the OPT dataset can achieve improvement over non BERT baselines.

Since the OPT dataset is composed of Twitter posts, we also assessed the performance of BERTweet (Nguyen et al., 2020). As BERTweet was solely pre-trained on tweets it has the potential to better interpret Twitter jargon. Indeed, we obtained an improvement of almost two percent by using BERTweet over BERT, on the OPT dataset with a mean accuracy of 84.58% on the validation set.

As a teacher model we chose the best BERTweet model on the validation set, out of our independent runs, having an accuracy of 85.01%. We can see in Table 1 the comparative results obtained for the knowledge distillation techniques. Even though the performance improvements are small we can conclude that by using PKD in combination with TA the system can learn to predict optimism at least as well as with a plain model.

Correlation between sentiment and optimism was studied by (Caragea et al., 2018). Their study revealed that there is no one-to-one correlation between sentiment polarity and optimism/pessimism. Still, common sense would suggest that optimism and pessimism should be revealed at least to some degree by the presence of certain feelings in a text.

We also studied the importance of the different distillation technique obtained by a student trained with distilled knowledge from multiple teachers. For our task, we trained two additional teachers, one on the Hate dataset and other on the Sent dataset. We used the best knowledge distillation technique with respect to the validation set, as shown in Table 1. Namely, we

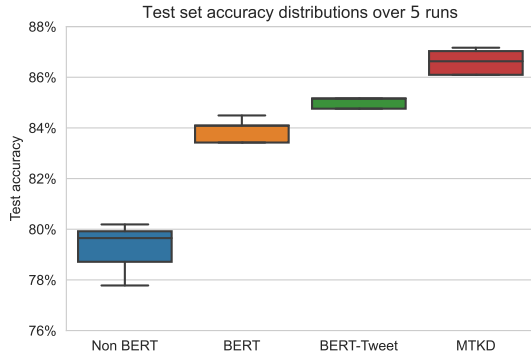


Figure 4: Accuracy distribution over five independent runs for optimism prediction. We can see that attention-based models BERT and BERTweet clearly outperform previous baselines. By using multi-task knowledge distillation we further improve models’ performance.

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 (Alshahrani et al., 2020)	84.25	–
BERT Base with SLA (Alshahrani et al., 2021)	85.69	–

Table 2: 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.

used PKD with TA. We selected the best student model based on its accuracy on the OPT validation set.

In Table 2 we can see the results obtained for multiple MTKD variations. We can see that both Hate and Sent teachers bring improvements in performance when used alongside an OPT teacher. Also, by using all three teachers we obtain the highest performance, of 86.60% on the test set.

We also performed an ablation study. First, the knowledge distillation technique proved to be of great importance. Without distillation the multi-task student is unstable and reveals a lower final accuracy of only 82.11%, highlighting the importance of a suitable distillation procedure. Second, the BERTweet model still needs to be used to obtain the best performance. By using vanilla BERT instead of BERTweet in the same MTKD setting we obtain an accuracy of only 85.64%. Third, the magnitude of the intermediate tasks is also relevant. If we downsample the Hate and Sent datasets to the dimension of the OPT dataset we obtain an accuracy of 86.19%.

A comparison between the performances of our models is displayed in Figure 4. We can see that the improvements revealed by our methods is consistent and

significant.

Finally, we also tested the best obtained models for the 1/ – 1 setting definition of the optimism. We show the results on Table 3. We can see that BERTweet and MTKD reveal the best results in this context as well clearly outperforming previous best models.

Model	Test Acc.	Dev Acc.
BERT	95.29	95.97
BERTweet	95.78	96.84
MTKD OPT + Hate + Sent	96.57	97.24
BERT Base with SLA (Alshahrani et al., 2021)	97.10	–
XLNet Base (Alshahrani et al., 2020)	96.16	–

Table 3: Model accuracies for predicting optimism in the setting 1/-1.

6. Error analysis

The OPT dataset is composed of tweets annotated for optimism on a scale from -3 to 3 . As it can be expected, the tweets with an annotation close to 0 may have a subtle presence of pessimism or optimism and this may very well be a subjective one. We noticed that our best model mostly misses mild annotated tweets, as can be seen in Figure 5. Most misses are tweets annotated in the interval $(0, 0.5]$.

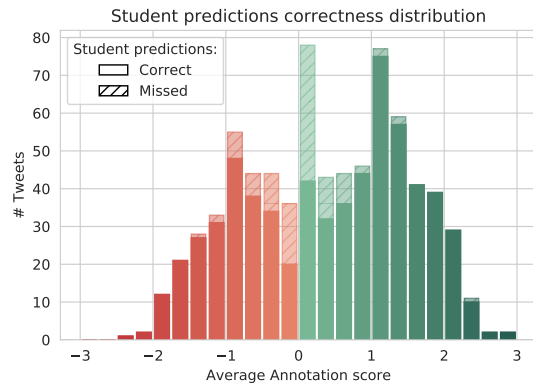


Figure 5: Distribution of the tweets correctly classified by the best model in comparison with the entire dataset. We can see that most missclassified tweets are in the annotation range $(-1, 1)$, where the differences between pessimism and optimism may be subjective.

In comparison with the best teacher model, the best student has a better performance on mild optimistic annotated tweets Figure 7. On the other hand, the teacher model still has a higher accuracy on mild pessimistic tweets. We came to the conclusion that this is the case since the Hate and Sent teachers were exposed to more positive tweets. Both Hate and Sent are unbalanced datasets containing significantly more positive and neutral tweets than negative ones, as also mentioned in section 4. This fact may have been the reason for which the student has not mastered the classification of the negative tweets. Therefore, a future direction to

	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%

Table 4: 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.

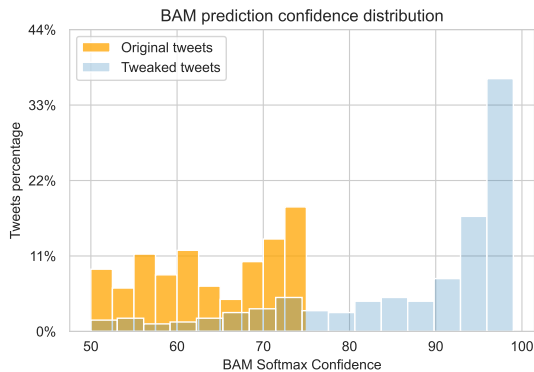


Figure 6: Softmax confidence distribution over the selected tweets before and after limited corrections. We can notice the prediction confidence increases significantly when tweets become more clear.

investigate would be the behavior of students that have both negative and positive teachers.

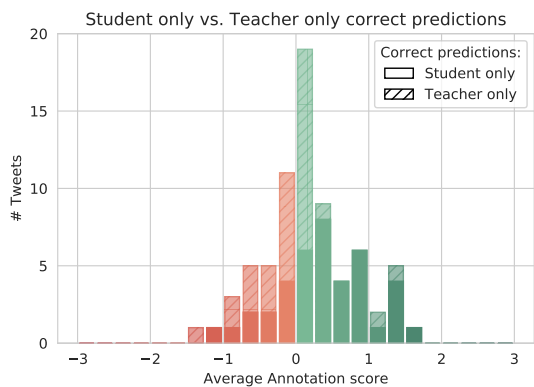


Figure 7: Distribution of tweets correctly classified by only the best teacher or the best student. We can see that the student outperforms its teacher on mildly optimistic tweets. The teacher has a better performance on the mildly pessimistic tweets.

Finally, we point to insights that could be gained from a qualitative analysis of "near misses" (i.e., predictions that are very close for the two classes of optimism and pessimism). Such an analysis suggests that the method developed may have trouble with statement in

which positive and negative words co-occur, especially in more complicated construction, such as interrogative sentences:

@if we don't raise taxes. how can we afford to pay the bills? you either cut spending or increase revenue (raise taxes).

More generally statements that alternate in one single construction positive and negative sentences, usually using one to illustrate the other, as we can see in:

the block feature on fb. i love it. if you bully me on my own page. you will be blocked. no discussion.", or "you are sometimes attracted to the shady side of love street w... more for scorpio.

A number of tweets are consistently missed by the models, and these may well be unbreakable hurdles in pushing the models to a higher accuracy. These tweets usually contain both positive and negative words in a melange that is often equally balanced, making it difficult or impossible even for the human observer to sort the tweet correctly. A good example of such a tweet is:

all is violent. all is bright.

This tweet contains one positive and one negative word and having the two equally balanced through the rest of the wording.

6.1. Model sensitivity to limited corrections

To better understand the near misses of our model we selected 51 correctly classified tweets and 40 misclassified tweets from the validation set. We were interested in tweets on which the model had a close output for both optimism and pessimism. Thus, we selected tweets classified with a softmax confidence lower than 75%.

We asked two psychology PhD students to independently modify each tweet in two ways to become first, more optimistic and second, more pessimistic. The adjustments were as limited as possible, in order to clarify the tweet. Thus, for each of the original selected tweet we obtained four new tweets. Two more optimistic and two more pessimistic. Each of the newly obtained tweets was annotated following the same procedure used to annotate the OPT dataset, described by

Ruan et al. (2016). Mainly, each of the modified tweets was rated on a pessimism - optimism scale between -2 and $+2$, by five independent annotators. We considered as new optimism gold standard the average of the five given ratings. Again, as discussed in section 4, we consider a tweet as pessimistic if the gold standard is smaller than or equal to 0 and optimistic otherwise. We show a tweet correction example in Table 4.

When applying our best student model (see Table 2) on the modified tweets we obtained an accuracy of 89.53%. Since only 56.04% of the selected tweets were correctly classified initially, we noticed that with minimal changes in the tweets, the accuracy of our model had a significant improvement.

We also observed that the softmax prediction confidence of the model increases (see Figure 6) and that most of the newly annotated tweets are classified with a softmax confidence over 75%.

7. Conclusion and Future Directions

In this study, we analyzed from a computational perspective the traits of optimism/pessimism and the feasibility of optimism/pessimism detection, based on the users' tweets. Towards this, we proposed a novel multi-task knowledge distillation framework to transfer knowledge from models trained on other complementary task (hate speech and sentiment analysis), such that the resulting model has the ability to discriminate between optimism and pessimism. Moreover, we also explored the linguistic characteristics that are expressed in optimism and pessimism and observed that the usage of some parts of speech directly correlate with the level of optimism. The results obtained for optimism in the current study are encouraging, not only for their practical application potential, but also because they open up avenues for future research. We suggest the need to further explore the relationship between optimism-pessimism and those mental health constructs they have been proven to be strongly connected to, such as depression. This will bring with it the supplementary challenge of studying reciprocal influences, as they unfold through time. The current study has also outlined the dire need to construct of a new, larger and better documented dataset for optimism/pessimism classification.

Acknowledgements

This work is partially supported by the NSF Grants IIS-2107487 and IIS-1912887. Any opinions, findings, and conclusions expressed here are those of the authors and do not necessarily reflect the views of NSF. This work was also partially supported by a grant of the Romanian Ministry of Education and Research, CC-CDI—UEFISCDI, project number 411PED/2020, code PN-III-P2-2.1-PED-2019-2271, within PNCDI III. The computation for this project was performed on Amazon Web Services through a research grant. We thank Raluca Dutu and Sergiu Condrea for providing the

tweets' limited corrections and for their valuable expertise. We would also like to thank our reviewers for their feedback and comments.

8. Bibliographical References

- Achat, H., Kawachi, I., Spiro, A., DeMolles, D. A., and Sparrow, D. (2000). Optimism and depression as predictors of physical and mental health functioning: the normative aging study. *Annals of Behavioral Medicine*, 22(2):127–130.
- Alarcon, G. M., Bowling, N. A., and Khazon, S. (2013). Great expectations: A meta-analytic examination of optimism and hope. *Personality and Individual Differences*, 54(7):821–827.
- Alshahrani, A., Ghaffari, M., Amirizirtol, K., and Liu, X. (2020). Identifying optimism and pessimism in twitter messages using xlnet and deep consensus. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.
- Alshahrani, A., Ghaffari, M., Amirizirtol, K., and Liu, X. (2021). Optimism/pessimism prediction of twitter messages and users using bert with soft label assignment. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.
- Caragea, C., Dinu, L. P., and Dumitru, B. (2018). Exploring optimism and pessimism in twitter using deep learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 652–658.
- Carver, C. S., Scheier, M. F., and Segerstrom, S. C. (2010). Optimism. *Clinical psychology review*, 30(7):879–889.
- Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.
- Clark, K., Luong, M.-T., Khandelwal, U., Manning, C. D., and Le, Q. V. (2019). BAM! born-again multi-task networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5931–5937, Florence, Italy, July. Association for Computational Linguistics.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Extremera, N., Durán, A., and Rey, L. (2009). The moderating effect of trait meta-mood and perceived stress on life satisfaction. *Personality and Individual Differences*, 47(2):116–121.
- Founta, A. M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Siriv-

- ianos, M., and Kourtellis, N. (2018). Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*.
- Gallagher, M. W., Long, L. J., and Phillips, C. A. (2020). Hope, optimism, self-efficacy, and posttraumatic stress disorder: A meta-analytic review of the protective effects of positive expectancies. *Journal of clinical psychology*, 76(3):329–355.
- Go, A., Bhayani, R., and Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009.
- Han, X. and Eisenstein, J. (2019). Unsupervised domain adaptation of contextualized embeddings for sequence labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4238–4248, Hong Kong, China, November. Association for Computational Linguistics.
- Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
- Kumar, U., Rana, V. S., Pykl, S., and Das, A. (2017). “A pessimist sees the difficulty in every opportunity; an optimist sees the opportunity in every difficulty” - understanding the psycho-sociological influences to it. In *Proceedings of the 14th International Conference on Natural Language Processing*.
- Mabry, E. A. (1974). Dimensions of profanity. *Psychological Reports*, 35(1):387–391.
- Martínez Alonso, H. and Plank, B. (2017). When is multitask learning effective? Semantic sequence prediction under varying data conditions. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 44–53, Valencia, Spain, April. Association for Computational Linguistics.
- McKenzie, C. R. and Nelson, J. D. (2003). What a speaker’s choice of frame reveals: Reference points, frame selection, and framing effects. *Psychonomic bulletin & review*, 10(3):596–602.
- Nguyen, D. Q., Vu, T., and Nguyen, A. T. (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*.
- Ortigosa, A., Martín, J. M., and Carro, R. M. (2014). Sentiment analysis in facebook and its application to e-learning. *Computers in human behavior*, 31:527–541.
- Owoputi, O., O’Connor, B., Dyer, C., Gimpel, K., Schneider, N., and Smith, N. A. (2013). Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 380–390.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037.
- Patrick, G. T. (1901). The psychology of profanity. *Psychological Review*, 8(2):113.
- Peterson, C. and Bossio, L. M. (2001). Optimism and physical well-being. *Optimism and pessimism: Implications for theory, research, and practice*, pages 127–145.
- Prati, G. and Pietrantoni, L. (2009). Optimism, social support, and coping strategies as factors contributing to posttraumatic growth: A meta-analysis. *Journal of loss and trauma*, 14(5):364–388.
- Pruksachatkun, Y., Phang, J., Liu, H., Htut, P. M., Zhang, X., Pang, R. Y., Vania, C., Kann, K., and Bowman, S. R. (2020). Intermediate-task transfer learning with pretrained language models: When and why does it work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5231–5247, Online, July. Association for Computational Linguistics.
- Qiu, B., Zhao, K., Mitra, P., Wu, D., Caragea, C., Yen, J., Greer, G. E., and Portier, K. (2011). Get online support, feel better—sentiment analysis and dynamics in an online cancer survivor community. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 274–281. IEEE.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.
- Rasmussen, H. N., Scheier, M. F., and Greenhouse, J. B. (2009). Optimism and physical health: A meta-analytic review. *Annals of behavioral medicine*, 37(3):239–256.
- Rosenthal, S., Farra, N., and Nakov, P. (2017). Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 502–518.
- Ruan, X., Wilson, S., and Mihalcea, R. (2016). Finding optimists and pessimists on twitter. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 320–325.
- Scheier, M. F. and Carver, C. S. (1992). Effects of optimism on psychological and physical well-being:

- Theoretical overview and empirical update. *Cognitive therapy and research*, 16(2):201–228.
- Scheier, M. F., Carver, C. S., and Bridges, M. W. (2001). Optimism, pessimism, and psychological well-being. *Optimism and pessimism: Implications for theory, research, and practice*, 1:189–216.
- Sun, S., Cheng, Y., Gan, Z., and Liu, J. (2019). Patient knowledge distillation for BERT model compression. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4323–4332, Hong Kong, China, November. Association for Computational Linguistics.
- Waterloo, S. F., Baumgartner, S. E., Peter, J., and Valkenburg, P. M. (2018). Norms of online expressions of emotion: Comparing facebook, twitter, instagram, and whatsapp. *new media & society*, 20(5):1813–1831.
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2019). Huggingface’s transformers: State-of-the-art natural language processing. *ArXiv*, pages arXiv–1910.
- Yan, L. and Tan, Y. (2014). Feeling blue? go online: an empirical study of social support among patients. *Information Systems Research*, 25(4):690–709.
- Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763.
- Zimmermann, J., Brockmeyer, T., Hunn, M., Schauenburg, H., and Wolf, M. (2017). First-person pronoun use in spoken language as a predictor of future depressive symptoms: Preliminary evidence from a clinical sample of depressed patients. *Clinical psychology & psychotherapy*, 24(2):384–391.