I am PsyAM: Modeling Happiness with Cognitive Appraisal Dimensions

Xuan Liu
Electrical Engineering and Computer Sciences
University of California Berkeley
USA
lxstephenlaw@gmail.com

Kokil Jaidka
Communications and New Media
National University of Singapore
Singapore
jaidka@nus.edu.sg

Abstract

This paper proposes and evaluates PsyAM[1], a framework that incorporates adaptor modules in a sequential multi-task learning setup to generate high-dimensional feature representations of hedonic well-being (momentary happiness) in terms of its psychological underpinnings. PsyAM models emotion in text through its cognitive antecedents, through auxiliary models that achieve multi-task learning through novel feature fusion methods. We show that BERT-PsyAM has cross-task validity and cross-domain generalizability through experiments with emotion-related tasks – on new emotion tasks and new datasets, as well as against traditional methods and BERT baselines. We further probe the robustness of BERT-PsyAM through feature ablation studies, as well as discuss the qualitative inferences we can draw regarding the effectiveness of the framework for representing emotional states. We close with a discussion of a future agenda of psychology-inspired neural network architectures.

1 Introduction

Governments are increasingly investing money into surveying and reporting nationwide psychological well-being as an indicator of success and wellness (Biswas-Diener et al., 2004), and some scholars have recommended monitoring social media for the unobtrusive measurement of regional trends in well-being and mental health. People are increasingly willing to post messages on social media to express their feelings. Words relate to emotions because they reflect how humans perceive their surroundings and ongoing events (Pennebaker et al., 2003); therefore, it is not surprising that language models trained on social media posts offer predictive insights into emotions. Emotions are an indicator of psychological states, such as *happiness* – the feelings of well-being related to momentary happiness or pleasure (Huta, 2016; Ryan and Deci, 2001).

Cognitive appraisal theory (CAT) posits that emotions result from how individuals appraise their situation and its impact on their well-being (Lazarus et al., 1980). CAT can model individual differences in emotional expressions; for instance, individuals may express happiness both, during a solitary walk as well as when they spend time with close friends if their core needs, drives, or motivations are suitably fulfilled. Therefore, we propose that relying on the stable cognitive antecedents of emotions could help us to train models that improve the state-of-the-art predictive accuracy for emotions detection tasks, and generalize more readily to other problems that infer psychological states from language, such as hedonic well-being.

In this work, we test a broad proposition that pretraining on cognitive auxiliary tasks improves emotion detection from text. This approach can bridge prior research in emotions classification with the increasing understanding of the link between self-expression and psychology. We offer the following contributions:

- New framework: We introduce **PsyAM** - a framework of **Psy**chological **A**dapter **M**odules[2] for emotion modeling that learns cognitive appraisal dimensions as auxiliary tasks that inform learning for a primary task.
- New tasks: We show that PsyAM (with BERT) improves over standard approaches in **cross-validation and replication** on new message-level data for predicting the duration of experienced happiness, and on user-level data for predicting the in-person emotional variance associated with their subjective well-being.
- Standard evaluation: We show that PsyAM (with BERT) outperforms the state-of-the-art

[1] https://github.com/stephenlaw30/BERT-PsyAM

[2] Code and data is at https://github.com/stephenlaw30/BERT-PsyAM

Findings of the Association for Computational Linguistics: ACL 2023, pages 1192–1210
July 9-14, 2023

in **cross-domain validation** on standard tasks for detecting emotion in binary- and multi-class settings.
- Real-world application: We demonstrate how PsyAM can offer real-world applications in predicting social media users' well-being through their Twitter posts.
- New annotated datasets: We have collected and annotated new happiness datasets from two nationally representative online surveys to validate our on new data and a real-world application.

Our work also offers a reality check to consider how leaderboard scores in emotion analysis translate to the real-world application of psycholinguistic models for unobtrusive mental health aggregates of communities. Recent studies have raised concerns about pre-trained psycholinguistic models' cross-domain and cross-task validity beyond simple emotion detection. Many word-trait relationships, such as that of self-referential language use and depression (Tackman et al., 2019), or the of social words and extraversion (Chen et al., 2020), inexplicably break down in different communicative contexts. We believe this is because of overfitting models to training corpora and the lack of psychologically motivated machine learning architectures. Prior work also often does not theorize *why* words relate to emotions or traits in the first place. Consequently, it has been challenging to build research depth in the representational modeling of the psychology of emotions and develop sophisticated neural network models. Herein lies an opportunity for social impact, as appropriately modeling and spatially aggregating emotions in social media make it possible to monitor mental health on a large scale (Dodds et al., 2015).

2 Background

Previous studies exploring the psycholinguistics of text have evaluated text classification approaches that can best correspond to the psychological measurements of different psychological traits and states (Turc et al., 2019; Guda et al., 2021; Guntuku et al., 2019). For instance, Buechel et al. (2018) elaborates on the psychological complexity of human reactions such as empathy and distress by annotating text data with the empathy assessments of their authors via multi-item scales and considering psychological tenets to be co-existent rather than correlated (Buechel et al., 2018).

In this study, we additionally offer two new tasks: predicting the *duration* of happiness and its *fluctuation* from linguistic expression. Both these concepts are grounded in psychology literature, where they are known to provide into emotional stability and well-being. Unlike a simple emotion detection task, the duration of happiness relates to distinguishing transient moods from persisting states of emotional well-being (Biswas-Diener et al., 2004). Therefore, it may require a more sophisticated representation of text semantics than one driven by affective words alone. Given that our primary datasets comprise textual descriptions of happy moments, the focus on hedonia is only appropriate. Secondly, emotional fluctuation has garnered a lot of recent interest as a more stable and tractable predictor of mental health as compared to mean-based emotion-based predictors, which are less sensitive to deviations and therefore to indicators of mental health issues. Emotional fluctuations expressed in verbal or written expressions has been evaluated for its relationship with within-person fluctuations in other personality states, including affective states (Sun et al., 2020; Golder and Macy, 2011), emotional experience (Back et al., 2010) and wellbeing (Pennebaker et al., 2003). In more recent work, emotional variance in social media posts has been also found to predict emotionally straining situations (Seraj et al., 2021), and daily emotional well-being (Lades et al., 2020).

We suggest that reorienting the understanding of emotions in text in its cognitive antecedents may offer a fruitful approach to its modeling, detection, and real-world application. We choose two cognitive constructs which, according to prior research, are often key to emotional appraisal (Ellsworth and Scherer, 2003; Karasawa, 1995; Moors, 2010). First, **Agency** reflects the role and accountability of an individual in a situation. Second, **Social Interaction** reflects the relationship of an individual with others in the situation (family, friends, or peers). Among many other appraisal dimensions, we find that agency and social interaction have also been explored in other perspectives on understanding how humans appraise happiness (Paulhus and Trapnell, 2008). These concepts were also applied to enrich the CLAff-HappyDB happiness dataset in Jaidka et al. (2020), but our work offers the first instance of applying them for the computational modeling of happiness. Furthermore, we extend the central premise of the authors to explore and model the

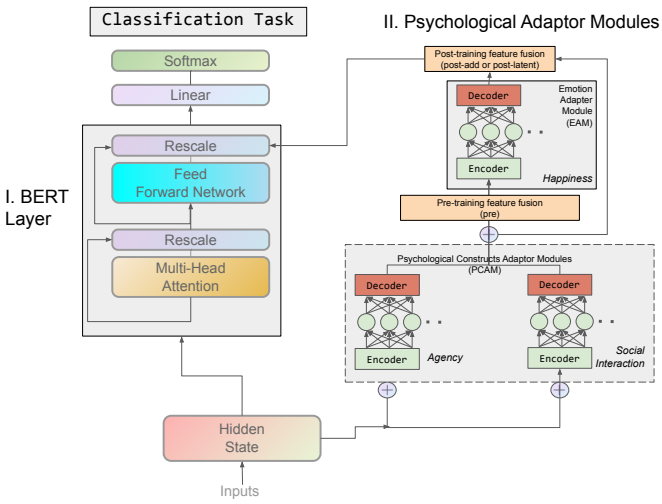


Figure 1: The BERT-PsyAM framework, demonstrating an input with two Psychological Adaptor Modules. The outputs are fused within the Emotion Adaptor Module and then passed to the next BERT layer.

duration of happiness.

Our choice of modeling the cognitive antecedents of emotions found no match in prior work.[3] Firstly, representing emotion at different semantic levels of abstraction, such as word embeddings (Bengio et al., 2000; Mikolov et al., 2013; Htait and Azzopardi, 2021), phrase- and sentence-level representations (Socher et al., 2013), or even paragraph-(Le and Mikolov, 2014) and document-level (Tang et al., 2015) representations involve evoked emotion rather than grounded emotion (Picard, 2000; Liu et al., 2017). Secondly, while finetuned LLMs like BERT offer state-of-the-art performance in emotion prediction, they are harder to adapt to multi-task learning problems. A finetuned BERT is also not ideal for knowledge transfer between multiple tasks. For instance, if we finetune BERT for n emotions, then we will need n times parameters inside BERT(n*110M), which would incur a huge memory overhead. Furthermore, our goal is to let different psychological tasks interact with the emotion task in the latent semantic space, which would necessitate re-engineering the end-to-end finetuning pipeline. These considerations made us opt for interactive adaptor modules with fewer parameters.

3 Methods

PsyAM is a broad proposition which can generalize to other frameworks. We chose to design our experiments centered on PsyAM coupled with BERT (BERT-PsyAM), which offers the best and most challenging opportunity to evaluate our claim in the high-dimensional space. This section defines and discusses our multi-task learning architecture (Figure 1), learning strategy, and the latent representations of appraisal dimensions.[4] In the rest of the paper, we have referred to the appraisal dimensions more generally as *psychological constructs*, and the adapter modules as *psychological adaptor modules* to suggest that the framework could generalize to include other kinds of cognitive and psychological antecedents of emotions.

3.1 BERT-PsyAM Architecture

In the BERT-PsyAM architecture for multi-task sequential learning of psychological constructs and emotion labels, we extended each BERT layer by attaching a Psychological Adaptor Module (PsyAM) in parallel, as seen in Figure 1.[5]

A PsyAM consists of trainable encoder-decoder weights along with multi-head attention and is instantiated uniquely for each psychological or emotional label. A PsyAM considers the activation state of the previous layer of BERT as its input. Further, the output of a PsyAM decoder is subject to recombination with the BERT layer as the residual connection and then serves as the input for the next BERT layer. The stack of PsyAMs is trained on the task of Psychological or Emotional Label Classification by attaching a unique classification head to the last layer of BERT-PsyAM.

PsyAM offers two innovations. First, for a

[3] We offer a more exhaustive discussion of related work in the Appendix.

[4] https://github.com/stephenlaw30/BERT-PsyAM

[5] A detailed description of all of the PsyAM components is reported in the Appendix.

deeper semantic understanding of the data, we designed **auxiliary tasks** to aid the predictive performance of the primary classification task in a sequential learning approach. Prior work offers a simplistic implementation of auxiliary pre-training (Mahmoud and Torki, 2020) through auxiliary contexts in a similarity detection task. Instead, we extend the auxiliary pre-training architecture with **multi-task learning** (Fifty, 2021), which aims to learn multiple different tasks simultaneously. Therefore, we are able to pre-train on auxiliary tasks that are not mutually exclusive, such as the cognitive appraisal dimensions underlying an emotional expression. Incorporating auxiliary pretraining has shown a performance improvement over the BERT models finetuned on individual tasks in prior work (Yu et al., 2019; Mahmoud and Torki, 2020), and we cororoborate these reports with similar findings in our paper.

We use **adaptor modules** for extracting latent semantic features from the text. The original idea of adaptor modules (Rebuffi et al., 2018) comes from the ResNet concept of residual connection in computer vision and it is a type of ensemble learning approach that can help overcome degradation problems in machine learning. The idea was adopted for natural language processing by Houlsby et al. (2019), who used an unfrozen encoder-decoder structure between the feedforward network and the Layer Norm operation while freezing the parameters of the original BERT layers, which has the effect of reducing the number of parameters to be finetuned but with a performance trade-off. Building on their idea,Stickland and Murray (2019) also incorporated attention mechanisms in adaptor modules. Herein lies our second innovation, as what we do is that in **fusing latent features**, we did not directly combine the encodings of text and numerical features as is typical in multimodal transformers, which could result in performance degradation due to dimensionality reduction. Instead, we generate a higher-dimensional feature representation with a layer-by-layer propagation rather than simply completing the splicing at the output and evaluate the effect of different fusion methods on the final predictive performance.

In summary, while the common way to use BERT for multiclass classification is to freeze the parameters of the original BERT and add several classification heads at the same level to fine-tune the parameters of the last layer to achieve multi-classification; however, because BERT's overall model parameters are fixed, finetuning in this manner often cannot achieve the best results, which would imply its inability to achieve optimal performance in every single classification task. Therefore, in the present architecture we assume that the internal parameters of BERT are not frozen. Furthermore, in BERT-PsyAM, we have replaced the original BERT Layer for fine-tuning with a small number of other parameters, so this is a great way for us to optimize processing with reduced overload.

4 Experiments

Our experiments address the question, *How well do the psychological constructs of agency and social interaction aid an understanding of emotions and well-being?* We evaluate the BERT-PsyAM framework on 5 different settings based on happiness, emotions, and well-being datasets. Through extensive experiments, we show how different variants of the BERT-PsyAM framework compares favorably to other approaches. With an ablation study and qualitative exploration, we reiterate the critical role of cognitive appraisal in well-being and emotion prediction tasks.

4.1 Task Settings and Datasets

4.1.1 Duration of Happiness task

Departing from typical emotion prediction tasks, we considered whether psychology labels could provide a deeper view into predicting the *duration of happiness*, a measure with immediate implications for understanding and modeling hedonic well-being (Biswas-Diener et al., 2004). The duration task was formulated as a binary classification problem that distinguished transient from more long-lasting feelings of happiness, on three datasets, with label distributions reported in Table 1:[6].

- CLAff-HappyDB: For training, testing, and internal validation, we relied on its 27,697 observations annotated with agency and social interaction labels (Jaidka et al., 2019).
- HappyDB-expand: The analysis was replicated on 59,664 further descriptions of happy moments (Asai et al., 2018) which constituted the superset of CLAff-HappyDB labeled through a semi-supervised approach.[7]

[6] 1 = "All day, I'm still feeling it" and 0 = "A few moments," "A few minutes, " "At least one hour," and "Half a day"

[7] Labeling for Agency and Social Interactions was done

- HappyDB-2021: A second replication was conducted on a freshly collected dataset of happy moments, sourced from a panel of internet users recruited through Qualtrics and then annotated through Amazon Mechanical Turk. The new dataset (N = 984) had a micro-level inter-annotator agreement of at least 80%. Dataset details are reported in the Appendix.

	CLAff-HappyDB	HappyDB-expand	HappyDB-2021
N	27,697	59664	984
Positive labels	10,807	20807	187
Agency	19,906	41233	796
Social interactions	15,369	34289	517

Table 1: Label distribution in the HappyDB datasets. We collected and annotated HappyDB-2021 for this study.

4.1.2 Emotion detection tasks

We evaluated the generalizability of PsyAM for *binary emotion classification* and *multi-category classification task*, on two popular industry benchmarks. Recent studies suggest that many affordances of Reddit make it imminently suitable for understanding and modeling individuals' physical and mental health (Wanchoo et al., 2023; Liu et al., 2023), including but not limited to self-disclosure and social interaction (Yang et al., 2017; Jaidka et al., 2019; Yang et al., 2019), both of which offer signals of agency and social interaction respectively.

- Kaggle SA-Emotions: We used the Kaggle SA-Emotions dataset[8] comprising 20,266 observations. We generated agency and social labels and subsequently predicted the Joy labels in a cross-validation setup.
- GoEmotions dataset: The GoEmotions dataset (Demszky et al., 2020) has finegrained emotion labels on 58k datapoints from Reddit, from which we sampled a balanced dataset of 14,589 datapoints with 7907 having a positive label of 'Joy'. We followed the same semi-supervised labeling and binary classification setup as the SA-Emotions task.

The label distributions for the two emotion datasets are reported in Table 2 and Figure 2.

using the best-performing BERT-PsyAM classifiers trained on CLAff-HappyDB. Classification accuracies are reported in the Appendix.

[8]https://www.kaggle.com/c/sa-emotions

	SA-Emotions	GoEmotions
Number of entries	20,266	14,589
Positive labels	5,209 (25.70%)	7,907 (54.2%)
Agency	10,907 (53.82%)	5311 (36.40%)
Social interaction	5,872 (28.97%)	4353 (29.84%)

Table 2: Binary label distribution in the Kaggle SA-Emotions and the GoEmotions datasets

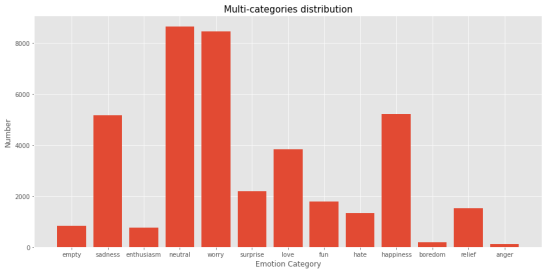


Figure 2: The distribution of labels in the SA-emotions.

4.1.3 Well-being prediction task

We also introduce the **TwitterUsers-2021** dataset to evaluate the real-world application of PsyAM to predict *user-level subjective-wellbeing*. The dataset comprises 217,910 tweets posted by 296 internet users recruited through Qualtrics, who took part in an online survey experiment and shared the link to their Twitter profiles.[9] We used the Twitter API to collect the social media posts of the 337 legitimate Twitter users who had participated in the survey and shared their Twitter handles. Of these, 296 participants had legitimate accounts and had posted at least one tweet of 50 characters or more that was not a retweet. As before, we used weakly supervised methods to generate Agency and Social Interaction labels for each post. Subsequently, we used BERT-PsyAM to generate labels for the duration of happy moments, which we used to derive an Emotional Fluctuation variable corresponding to the within-person variance in the happiness scores.

4.2 Experiments

We performed the following experiments to test and validate the BERT-PsyAM framework:

- **Cross-validation:** The CLAff-HappyDB dataset with an 80-10-10 split is used for internal validation for the duration task.
- **Replication analysis:** Experiments are replicated on the HappyDB-expand dataset and the HappyDB-2021 datasets, which constitute held-out data and data from a different cultural context.

[9]Participant demographics are reported in the Appendix.

Table 3: Predictive performance on the duration task, sorted by accuracy on the CLAff-HappyDB dataset. The color gradient identifies the best performing models for each metric (darker is better). * shows that the best result is significantly better than the best BERT baseline ($p < 0.05$).

	Dataset	CLAff-HappyDB			HappyDB-expand			HappyDB-2021		
		ACC	ROC	F-1	ACC	ROC	F-1	ACC	ROC	F-1
Traditional Method	GaussianNB	56.06	59.97	56.95	56.76	60.01	54.20	54.90	59.31	55.90
	MLPclassifier	63.72	61.39	51.38	65.76	62.20	50.42	63.96	61.12	49.51
	MLPclassifier+A+S	63.75	61.61	52.42	64.54	61.42	50.82	64.13	62.32	54.45
BERT-based Method	BERT-FT	70.22	75.46	59.74	71.63	76.08	57.71	71.07	75.69	59.94
	BERT-PALS	70.87	75.64	60.34	71.54	76.19	57.66	71.31	76.83	58.95
Proposed Method	BERT-PsyAM *post-add*	71.44	77.10	61.92	72.71	77.30	59.75	71.87	77.27	61.70
	BERT-PsyAM *pre*	71.59	77.03	62.68	72.61	77.51	60.75	71.98	76.78	62.11
	BERT-PsyAM *post-linear*	72.13*	77.41*	62.52*	72.75*	77.47*	60.17*	72.08*	77.27*	62.52*

Table 4: Predictive performance on emotion detection, sorted by accuracy on the binary task. The color gradient identifies the best performing models for each metric (darker is better). * $p < 0.05$.

Approach		Kaggle SA-Emotion			
		Binary			Multi-class
		ACC	ROC	F-1	ACC
Traditional Method	GaussianNB	75.09	68.35	55.11	18.58
	MLPclassifier	78.29	71.73	56.35	26.93
	MLPclassifier+A+S	79.13	73.01	60.72	27.72
BERT-based Method	BERT-FT	84.16	87.28	67.21	35.32
	BERT-PALS	84.51	89.32	68.91	36.72
Proposed Method	BERT-PsyAM *post-linear*	85.15	89.64	71.63	39.84
	BERT-PsyAM *post-add*	85.29	90.23	71.29	39.48
	BERT-PsyAM *pre*	85.35*	89.90*	71.90*	39.39*

Table 5: Predictive performance on GoEmotions. The color gradient identifies the best performing models for each metric (darker is better). * $p < 0.05$.

Approach		GoEmotions		
		ACC	ROC	F-1
Traditional Method	GaussianNB	73.00	74.57	71.41
	MLPclassifier	88.76	88.67	89.45
	MLPclassifier+A+S	89.10	89.10	89.97
BERT-based Method	BERT-FT	91.30	96.98	91.77
	BERT-PALS	91.09	96.64	91.76
Proposed Method	BERT-PsyAM *post-linear*	93.69*	98.08*	94.12*
	BERT-PsyAM *post-add*	93.56	97.86	94.03
	BERT-PsyAM *pre*	92.39	97.84	92.90

- **External validation:** We used the Kaggle SA-Emotions and the GoEmotions datasets to benchmark the external validity of BERT-PsyAM with an 80-10-10 data split for binary and multi-class emotion prediction.
- **Ablation analysis:** We conducted an ablation experiment to evaluate the role of different combinations of agency and social interaction labels in predictive performance. We also included other contextual information (e.g. the reflection period) and respondent demographics (their marital status, which would affect their psychological outlook (Diener et al., 2000)).[10]
- **Model visualization:** We applied the Captum toolkit (Kokhlikyan et al., 2020) to visualize the impact of individual words on the classification confidence score for the duration task. Captum calculates the layer integrated gradient on the test cases input through BERT-PsyAM with *post-linear* feature fusion. We then compared word attributions from token embeddings by a BERT-finetuned vs. a BERT-PsyAM model.
- **Real-world application:** We evaluated the predictive performance of Bert-PsyAM on user-level subjective well-being prediction based on duration of happiness scores generated on their Twitter posts.

4.3 Model Setup

We adapted the PyTorch *BERT-base-uncased* as our backbone, initialized with pre-trained weights. Our implementation comprises 12 layers, 12 attention heads and 768 as the hidden unit size. We modeled psychological constructs (agency, social) prediction as a binary classification task. We set the augmentation size a to be 204. We trained for each task with the AdamW optimizer (Loshchilov and Hutter, 2019) for 3 epochs, with a batch size of 32, a learning rate of 3e-5, and a maximum sequence length of 128 tokens.

We have compared the BERT-PsyAM framework against traditional and BERT-based methods.

[10]These experiments were conducted on the primary task and dataset with *post-linear* feature fusion.

Traditional methods adopt the pre-trained word embeddings (Mikolov et al., 2018) and the bag-of-words model to generate the sequence representation which is passed to classifiers: Gaussian Naive Bayes (Gaussian-NB) and Multi-layer Perceptron classifier (MLP-classifier). In order to compare role of different representation of psychological features, we made a baseline called MLP-classifier+A+S which embeds agency and social information into the input of MLP-classifier with numerical features. Since BERT-PsyAM is constructed with a BERT backbone, BERT-Finetune and BERT-NPALS (Stickland and Murray, 2019) offer the ideal baselines to illustrate performance improvements.

5 Results & Analysis

As seen in Table 3, first, in the **internal cross-validation**, BERT-PsyAM with the linear post-training fusion strategy is the best performed model in the internal validation by surpassing baselines 1.5% in accuracy and nearly 2% in both ROC and F-1 in repeated iterations (Accuracy = 72.13 vs. 70.87; ROC = 77.41 vs. 75.64, $p < 0.05$). Next, in the **replication analysis** reported for HappyDB-expand, we see similar improvements as above (Accuracy = 72.75 vs. 71.54; ROC = 77.47 vs 76.19, $p < 0.05$), illustrating its robustness and general ability that leverages pre-trained models. Finally, the performance on HappyDB-2021 (Accuracy = 72.08 vs. 71.31; ROC = 77.27 vs 76.83, $p < 0.05$) suggests that the model also generalizes well to new cultural contexts.

5.1 External validation

Tables 4 and 5 and Table shows PsyAM's performance at emotion classification. For SA-Emotions, in the binary classification task, BERT-PsyAM with pre-training feature fusion has the highest Accuracy and ROC that are statistically significantly higher than BERT-PALS by about 1% each in repeated runs (Accuracy = 85.35 vs 84.51; ROC = 89.90 vs. 89.32, $p < 0.05$), while for GoEmotions the post-linear feature fusion has the better performance. BERT-PsyAM also outperforms BERT-PALS in multi-task prediction by about 3% (Accuracy = 71.90 vs 68.91; ROC = 39.39 vs. 36.72, $p < 0.05$), suggesting that our framework has generalized well to new tasks, datasets, and emotions.

As seen in Table 6, adding both psychological constructs considerably improves over BERT-finetune (ROC = 77.41 vs. 75.64). We also found that the performance increased further with the addition of a **third adapter module** related to emotional transience (Asai et al., 2018) - the reflection period (ROC = 77.67 vs 75.64).

5.2 Model visualization

Figure 3 reports a confusion matrix of classifications by BERT-FT reported in Table 3 vs. BERT-PsyAM on CLAff-HappyDB data. Across the four quadrants, BERT-FT does attribute importance to emotion words such as *'anniversary'* and *'special,'* but also to irrelevant words such as *'to,'* and *'it's'*, confirming recent findings regarding BERT (Hayati et al., 2021). In contrast, BERT-PsyAM generates higher confidence scores and appears to pay attention to first person and possessive pronouns that denote agency (Rouhizadeh et al., 2018) (*'me'* and *'my'*), and social relationships and interactions (Jaidka et al., 2020) (*'husband'* and *'daughter'*; *'we'* and *'our'*).

5.3 Real-world application

In Figure 4 (a), we demonstrate whether labels generated using PsyAM offer any psychological insights for real-world applications. In a regression analysis that predicts subjective well-being, after controlling for demographic covariates, higher emotional fluctuation scores based on BERT-PsyAM predictions predict lower wellbeing, a relationship which corroborates prior work in psychology (Seraj et al., 2021). The value of the coefficients suggests that a 1% decrease in emotional fluctuation predicts a 0.04% increase in the subjective well-being of an individual, or a magnitude of 0.4 points on a 10-point scale. Although the predicted R^2 values are low, this is typical of models that use linguistic covariates to predict psychological traits (Boyd and Pennebaker, 2017). In Figure 4(b), We examined the interactive effect of age on this relationship and find that the interactive effect of age and emotional fluctuation is significant (and negative) among 18-34 year olds.

6 Discussion

Our findings offer the following insights towards a future agenda of neural networks inspired by human psychology:

- Feature-rich representation: PsyAM appears to help BERT pay more attention to words reflecting cognitive constructs, such as words

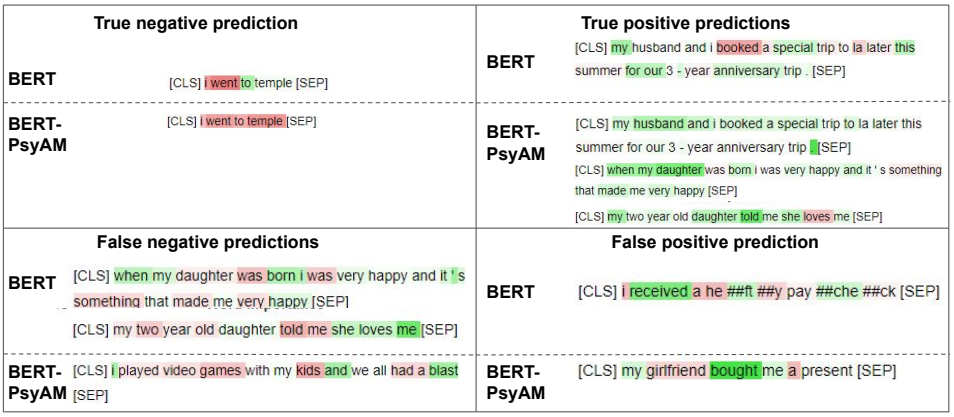


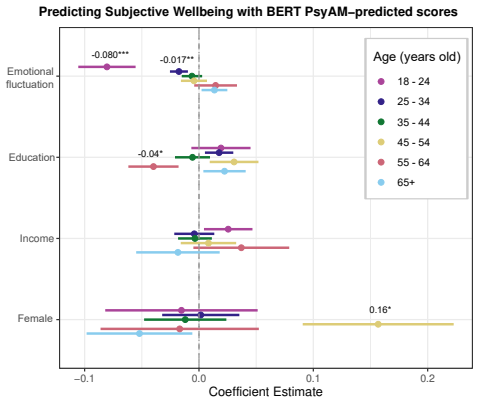
Figure 3: A Captum-based visualization of the accurate and inaccurate classifications by the best-performing BERT baseline and the BERT-PsyAM framework with *post-linear* feature fusion. Darker (lighter) shades of green (red) indicates a higher (lower) magnitude of importance (penalty). Best seen in color.

Predicting Subjective Wellbeing: Model Summary

Coefficient	Estimate (b)	Std. Error	p (* if < 0.05)
(Intercept)	0.559	0.084	***
Emotional fluctuation	-0.038	0.011	***
Education	0.006	0.006	
Income	-0.001	0.008	
Female	0.009	0.018	
Age	-0.048	0.022	*
Emotional fluctuation x Age	0.009	0.003	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(a)



(b)

Figure 4: (a) The model summary of a regression model predicting subjective well-being as a function of user demographics and the emotional fluctuation (predicted by Bert-PsyAM) reflected on Twitter (N=296). (b) Predicting subjective wellbeing with BERT PsyAM-predicted scores for emotional fluctuation, on the TwitterUsers2021 dataset. (*: $p < 0.05$)

Table 6: Ablation study on the duration task with CLAff-HappyDB. The color gradient identifies the best performing models for each metric (darker is better).

	BERT	1	2	3	4
Agency		✓		✓	✓
Social interaction			✓	✓	✓
Reflection period					✓
ACC	70.87	70.82	71.58	72.13	71.94
ROC	75.64	77.12	77.36	77.41	77.67
F-1	60.34	61.49	62.82	62.52	63.09

referencing the self, family members, and social activities, as seen in Figure 3.

- Operational flexibility: Within PsyAM, different feature fusion methods had the best performance, suggesting that contextual and dataset differences can be accounted for through minor adjustments of the PsyAM framework.
- Robust predictions: PsyAM works well with a variety of datasets, cultural contexts, and even tasks. Even when the label distribution differs from the original, the superior performance of PsyAM models signals that our architecture is robust to new data.
- Semi-supervised extensions: Labeling existing datasets with models finetuned on small annotated datasets offers a low-cost alternative to obtaining high-quality annotations. This is

the approach we used, as reported in Table 8 of the Appendix, and the predictions have face-validity, as reported in Figure 3.

- Cognitive construct extensions: Other cognitive constructs, such as reflection period, can offer additional measures of cognitive complexity to further improve the modeling of emotional expression.

7 Conclusion

The main advantage of the BERT-PsyAM framework is a modeling paradigm that transfers to new domains and tasks for detecting emotions and, by extension, well-being. Psychologists can use PsyAM to build and test new hypotheses about self-expression, cognitive appraisal, and behavior. It could also be helpful in interventionist scenarios requiring live monitoring and reporting problematic social media posts.

PsyAM achieved substantive improvements over the state-of-the-art BERT alternatives, and we show that this is because of the contribution of the high-dimensional feature representation inside the Adaptor Modules. Different feature fusion methods achieved different degrees of improvement in different settings, and different latent variables may add value in different problem contexts.

Future directions: We have released the labeled datasets developed as part of the study for researchers to explore further how psychological traits and states can inspire better neural architectures for text classification. We plan to explore other appraisal dimensions, such as goal conduciveness and certainty, as well as individual differences through demographic and personality traits.

Limitations: The duration task focused only on data with a positive happiness label, but it would be interesting to see whether the framework generalizes to a complete dataset and more sophisticated problem definitions. The need for annotations limits the generalizability of our approach, but the BERT-PsyAM framework is effective even with labels generated through semi-supervised methods and other metadata.

Ethical considerations: The models are intended for aggregate- and group-level inferences, and not individual or message-level inferences. Despite our cross-domain validation efforts, we caution that relying exclusively on AI-inferred relationships between emotion, self-efficacy, and self-determination may lead to inaccurate measurements. Finally, models trained in a specific socio-cultural setting may nevertheless violate the social conventions in specific settings, such as in the workplace, and cultural conventions of individualism and collectivism in social life (Diener et al., 2009).

Acknowledgments: We thank Niyati Chhaya, Chaitanya Aggarwal, and Gerard Yeo for feedback on early versions of this work. This work was supported by an NUS CTIC grant and a Nanyang Presidential Postdoctoral fellowship.

References

Akari Asai, Sara Evensen, Behzad Golshan, Alon Halevy, Vivian Li, Andrei Lopatenko, Daniela Stepanov, Yoshihiko Suhara, Wang-Chiew Tan, and Yinzhan Xu. 2018. Happydb: A corpus of 100,000 crowdsourced happy moments. *arXiv preprint arXiv:1801.07746*.

Mitja D Back, Albrecht CP Küfner, and Boris Egloff. 2010. The emotional timeline of september 11, 2001. *Psychological Science*, 21(10):1417–1419.

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. *Advances in Neural Information Processing Systems*, 13.

Robert Biswas-Diener, Ed Diener, and Maya Tamir. 2004. The psychology of subjective well-being. *Daedalus*, 133(2):18–25.

Ryan L Boyd and James W Pennebaker. 2017. Language-based personality: a new approach to personality in a digital world. *Current opinion in behavioral sciences*, 18:63–68.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4758–4765.

Jiayu Chen, Lin Qiu, and Moon-Ho Ringo Ho. 2020. A meta-analysis of linguistic markers of extraversion: Positive emotion and social process words. *Journal of Research in Personality*, 89:104035.

Edward L Deci and Richard M Ryan. 2000. The" what" and" why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological inquiry*, 11(4):227–268.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ed Diener, Marissa Diener, and Carol Diener. 2009. Factors predicting the subjective well-being of nations. In *Culture and well-being*, pages 43–70. Springer.

Ed Diener, Carol L Gohm, Eunkook Suh, and Shigehiro Oishi. 2000. Similarity of the relations between marital status and subjective well-being across cultures. *Journal of cross-cultural psychology*, 31(4):419–436.

Peter Sheridan Dodds, Eric M Clark, Suma Desu, Morgan R Frank, Andrew J Reagan, Jake Ryland Williams, Lewis Mitchell, Kameron Decker Harris, Isabel M Kloumann, James P Bagrow, et al. 2015. Human language reveals a universal positivity bias. *Proceedings of the national academy of sciences*, 112(8):2389–2394.

Phoebe C Ellsworth and Klaus R Scherer. 2003. *Appraisal processes in emotion.* Oxford University Press.

Nicholas Epley and Juliana Schroeder. 2014. Mistakenly seeking solitude. *Journal of Experimental Psychology: General*, 143(5):1980.

Christopher Fifty. 2021. Deciding which tasks should train together in multi-task neural networks.

Scott A Golder and Michael W Macy. 2011. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, 333(6051):1878–1881.

Bhanu Prakash Reddy Guda, Aparna Garimella, and Niyati Chhaya. 2021. EmpathBERT: A BERT-based framework for demographic-aware empathy prediction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3072–3079, Online. Association for Computational Linguistics.

Sharath Chandra Guntuku, Anneke Buffone, Kokil Jaidka, Johannes C Eichstaedt, and Lyle H Ungar. 2019. Understanding and measuring psychological stress using social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 214–225.

Shirley Hayati, Dongyeop Kang, and Lyle Ungar. 2021. Does bert learn as humans perceive? understanding linguistic styles through lexica. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6323–6331.

John F Helliwell and Robert D Putnam. 2004. The social context of well-being. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 359(1449):1435.

Dan Hendrycks and Kevin Gimpel. 2016. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. *CoRR*, abs/1606.08415.

Cynthia A Hoffner and Sangmi Lee. 2015. Mobile phone use, emotion regulation, and well-being. *Cyberpsychology, Behavior, and Social Networking*, 18(7):411–416.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Amal Htait and Leif Azzopardi. 2021. Awessome: An unsupervised sentiment intensity scoring framework using neural word embeddings. In *European Conference on Information Retrieval*, pages 509–513. Springer.

Veronika Huta. 2016. An overview of hedonic and eudaimonic well-being concepts. *The Routledge handbook of media use and well-being*, pages 14–33.

Kokil Jaidka, Niyati Chhaya, Saran Mumick, Matthew Killingsworth, Alon Halevy, and Lyle Ungar. 2020. Beyond positive emotion: Deconstructing happy moments based on writing prompts. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 294–302.

Kokil Jaidka, Saran Mumick, Niyati Chhaya, and Lyle Ungar. 2019. The cl-aff happiness shared task: Results and key insights. In *AffCon@ AAAI*.

Kaori Karasawa. 1995. Cognitive antecedents of emotions findings and future directions. *Japanese Psychological Research*, 37(1):40–55.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. 2020. Captum: A unified and generic model interpretability library for pytorch.

Margie E Lachman and Suzanne L Weaver. 1998. The sense of control as a moderator of social class differences in health and well-being. *Journal of personality and social psychology*, 74(3):763.

Leonhard K Lades, Kate Laffan, Michael Daly, and Liam Delaney. 2020. Daily emotional well-being during the covid-19 pandemic. *British journal of health psychology*, 25(4):902–911.

Richard S Lazarus, Allen D Kanner, and Susan Folkman. 1980. Emotions: A cognitive–phenomenological analysis. In *Theories of emotion*, pages 189–217. Elsevier.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR.

Tingting Liu, Devansh Jain, Shivani Reddy Rapole, Brenda Curtis, Johannes C Eichstaedt, Lyle H Ungar, and Sharath Chandra. 2023. Detecting symptoms of depression on reddit. In *Proceedings of the 15th ACM Web Science Conference 2023*, pages 174–183.

Vicki Liu, Carmen Banea, and Rada Mihalcea. 2017. Grounded emotions. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 477–483. IEEE.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.

Mufan Luo and Jeffrey T Hancock. 2020. Self-disclosure and social media: motivations, mechanisms and psychological well-being. *Current Opinion in Psychology*, 31:110–115.

Somaia Mahmoud and Marwan Torki. 2020. Alexu-auxbert at semeval-2020 task 3: Improving bert contextual similarity using multiple auxiliary contexts. In *Proceedings of the 14th International Workshop on Semantic Evaluation*.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751.

Agnes Moors. 2010. *Theories of emotion causation: A review*. Psychology Press.

Becky Lynn Omdahl. 2014. *Cognitive appraisal, emotion, and empathy*. Psychology Press.

Delroy L Paulhus and Paul D Trapnell. 2008. Self-presentation of personality: An agency-communion framework.

James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, 54(1):547–577.

Rosalind W Picard. 2000. *Affective computing*. MIT press.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Niyati Chhaya, et al. 2021. Recognizing emotion cause in conversations. *Cognitive Computation*, 13(5):1317–1332.

Sylvestre-Alvise Rebuffi, Andrea Vedaldi, and Hakan Bilen. 2018. Efficient parametrization of multi-domain deep neural networks. pages 8119–8127.

Masoud Rouhizadeh, Kokil Jaidka, Laura Smith, H. Andrew Schwartz, Anneke Buffone, and Lyle Ungar. 2018. Identifying locus of control in social media language. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Richard M Ryan and Edward L Deci. 2001. On happiness and human potentials: A review of research on hedonic and eudaimonic well-being. *Annual review of psychology*, 52:141.

Gillian M Sandstrom and Elizabeth W Dunn. 2014. Social interactions and well-being: The surprising power of weak ties. *Personality and Social Psychology Bulletin*, 40(7):910–922.

Sarah Seraj, Kate G Blackburn, and James W Pennebaker. 2021. Language left behind on social media exposes the emotional and cognitive costs of a romantic breakup. *Proceedings of the National Academy of Sciences*, 118(7):e2017154118.

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*, pages 1631–1642.

Asa Cooper Stickland and Iain Murray. 2019. BERT and PALs: Projected attention layers for efficient adaptation in multi-task learning. *CoRR*, abs/1902.02671.

Jessie Sun, H Andrew Schwartz, Youngseo Son, Margaret L Kern, and Simine Vazire. 2020. The language of well-being: Tracking fluctuations in emotion experience through everyday speech. *Journal of Personality and Social Psychology*, 118(2):364.

Allison M Tackman, David A Sbarra, Angela L Carey, M Brent Donnellan, Andrea B Horn, Nicholas S Holtzman, To'Meisha S Edwards, James W Pennebaker, and Matthias R Mehl. 2019. Depression, negative emotionality, and self-referential language: A multi-lab, multi-measure, and multi-language-task research synthesis. *Journal of personality and social psychology*, 116(5):817.

Duyu Tang, Bing Qin, and Ting Liu. 2015. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1422–1432.

Louis Tay and Ed Diener. 2011. Needs and subjective well-being around the world. *Journal of personality and social psychology*, 101(2):354.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962v2*.

Karan Wanchoo, Matthew Abrams, Raina M Merchant, Lyle Ungar, and Sharath Chandra Guntuku. 2023. Reddit language indicates changes associated with diet, physical activity, substance use, and smoking during covid-19. *Plos one*, 18(2):e0280337.

Diyi Yang, Zheng Yao, and Robert Kraut. 2017. Self-disclosure and channel difference in online health support groups. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 704–707.

Diyi Yang, Zheng Yao, Joseph Seering, and Robert Kraut. 2019. 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*, pages 1–15.

Lele Yu, Shaowu Zhang, Yijia Zhang, Hongfei Lin, et al. 2021. Improving human happiness analysis based on transfer learning: Algorithm development and validation. *JMIR Medical Informatics*, 9(8):e28292.

Shanshan Yu, Jindian Su, and Da Luo. 2019. Improving bert-based text classification with auxiliary sentence and domain knowledge. *IEEE Access*, 7:176600–176612.

A Appendix

A.1 Background

Psychological constructs including but not limited to empathy, distress, agency, social affinity have been widely studied by computational linguists (Turc et al., 2019; Guda et al., 2021; Rouhizadeh et al., 2018; Guntuku et al., 2019). Such studies, primarily, evaluate representational learning approaches for downstream tasks to predict the appropriate psychological constructs associated with the given text. For instance, Buechel et al. (2018) elaborates on the psychological complexity of human reactions such as empathy and distress, by annotating text data with the empathy assessments of their authors via multi-item scales. Even this study considered empathy and distress to be co-existent rather than correlated, offering little insight into the role of cognitive appraisals as the wellspring of emotional expression (Omdahl, 2014; Hoffner and Lee, 2015). We formulate our problem as the multi-task learning of psychological states in tandem with human emotions, which offers insight into the role of psychological self-appraisals as the wellspring of emotional expression (Luo and Hancock, 2020; Omdahl, 2014; Hoffner and Lee, 2015). In the following paragraphs, we first motivate the problem of modeling emotion in terms of its psychological underpinnings, followed by a review of the literature on multitask learning and language modeling that is relevant to our context.

A.2 State of the art in emotion prediction

Emotion in text has been modeled in different studies at different levels of abstraction and meaning, such as sentiment-specific word embeddings (Bengio et al., 2000; Mikolov et al., 2013; Htait and Azzopardi, 2021), phrase- and sentence-level representations (Socher et al., 2013), and even at the paragraph-(Le and Mikolov, 2014) and document-level (Tang et al., 2015). Note that these approaches appear to conflate the detection of expressed versus evoked emotion (Picard, 2000), as they tend further and further away from the root of emotion (Liu et al., 2017). We suggest that grounding the emotion prediction task in its cognitive antecedents, similar to prior work focusing on its causes (Poria et al., 2021), may offer a fruitful approach to its modeling and detection.

In studies with a temporal component, Long-Short Term Memory (LSTM) architectures and Gated Neural Networks have been used successfully; however, they often fall short of transformer-based models in tasks involving language understanding, such as emotion prediction. Therefore, given the stability and universality of transformer-based models, and the ability to benchmark against previous work and datasets, we opted to use the BERT model as the backbone for our experiments. BERT (Devlin et al., 2018) is a pre-training language model with two self- supervised pretraining objectives: masked language modeling (MLM) and next sentence prediction (NSP). The NSP task in pre-training is used to fine-tune the model using labeled data in order to obtain the prediction result. Although a finetuned BERT model is known to achieve good results in emotion prediction, it is not suitable for our purpose for two reasons. Firstly, a finetuned BERT is not ideal for knowledge transfer between multiple tasks. For instance, if we finetune BERT for n emotions, then we will need n*110M parameters, which would incur a huge memory overhead. Secondly, our goal is to let different psychological tasks interact with the emotion task at the semantic latent space, which implied that finetuned models would not meet our requirements. These considerations made us opt for an interactive module with fewer parameters.

A.3 Multi-task learning

A multi-task learning setup aims to learn multiple different tasks simultaneously. This can be thought of as predicting attributes of text that are not mutually exclusive, such as the psychological and emotional facets of the memory of a happy moment. In multi-task learning, more than one loss function is trained, or a part of the loss function comes from a different task, with the expectation that the model will apply the information it learns during training on one task to decrease the loss on other tasks included in training the network (Fifty, 2021). The initial proposal of adaptor modules for natural language processing by Houlsby (Houlsby et al., 2019) offered an end-to-end structure which offers fine-tuning parameters while employing fewer parameters than the entire BERT, but with a degraded final performance. However, their final performance was slightly worse than the whole BERT finetune model while reducing the number of parameters that must be fine-tuned.

The BERT-PsyAM architecture proposed in this study adapts the BERT and PALS architecture (Stickland and Murray, 2019) which included projected attention layers, as parts within the adaptor module. This constitutes an ensemble approach which is then used for the sequential training of psychological states and emotions. Our approach is novel in three ways: firstly, instead of using adaptors for parameter minimization, we use them to extract latent semantic features with the multi-head attention mechanism and feedforward layers. Secondly, instead of adding attention outputs, we use residual connections to combine the latent semantic features between tasks. This allows different tasks to interact and support each other with high-dimensional (context-psychology) feature representations. In this paper, we have evaluated three different methods for feature fusion, all of which perform well. Thirdly, in the process of fusing latent features, we did not directly combine the encodings of text and numerical features as is typical in multimodal transformers, which could result in signal losses due to dimensionality reduction. Instead, we generate a higher-dimensional (context & psychology) feature representation with a layer-by-layer propagation rather than simply completing the splicing at output.

A.4 Context

Now that we have motivated our architecture, we will explain the theoretical concepts it realizes. This study focuses on the psychological concepts that are meaningful in the understanding of emotion – specifically, in understanding happiness. Psychologists define *agency* as the feeling of being in control of one's life. It is related to the ideas of autonomy (Tay and Diener, 2011), and self efficacy (Deci and Ryan, 2000), which are known to have a strong relationship with personal health and well-being (Lachman and Weaver, 1998). Moreover, its linguistic correlates have been examined in prior work (Rouhizadeh et al., 2018). On the other hand, *social interaction* and feeling connected to others are also central to well-being (Helliwell and Putnam, 2004), the feeling of belongingness (Sandstrom and Dunn, 2014), and happiness (Epley and Schroeder, 2014).

The association of agency and social interaction in textual descriptions of happy moments was explored in the The CLAff-HappyDB Shared Task models happy moments in terms of their agency and social interaction (Jaidka et al., 2019). The Task proposes that a happy moment may involve either agency and social interaction, both, or neither of these; however, their interplay could be helpful for an enriched understanding of happiness. However, the labeled dataset was small in size, limiting the possibilities for technical and conceptual follow-up work. In this study, we have used semi-supervised approaches to expand on the training data, collected new data, and tested our approach on standard emotion detection tasks on a well-known emotion dataset.

B BERT-PsyAM architecture

B.0.1 BERT Layer

Figure 1 illustrates how a BERT layer receives the activation state of the previous BERT layer as input, and generates d_h dimensional hidden states for each token in the sequence, which are forwarded to the BERT attention and the feedforward network.

Multi-Head Attention: N different attention heads[11] are applied to extract the attention score on tokens as the aggregate of previous hidden states.

Rescale: This sub-module splices and re-scales the feature space to make it more suitable for the task in context. It does so by calculating the mean

[11] We have used N = 12

μ and variance σ of each unit in the input vector x, and then re-scaling them by learning a gain *g* and bias *b* parameter sequentially.

Feed-forward Networks (FFN) Next, an FFN framework containing two linear transformations with a GELU activation function (Hendrycks and Gimpel, 2016) is applied to each position of hidden states by transforming them to a high-dimensional space and then transforming them back. Here we add BA_N, the residual part we got from the multi-head attention setup, which gives us the final BERT-PsyAM layer output as:

$$Layer\_output = Rescale(FFN(BA_N(x)) + BA_N(x) + PsyAM(x)) \quad (1)$$

The PsyAM(x) component of Equation 1 is explained in the following paragraphs.

B.0.2 Adaptor Modules

There are two types of Adaptor Modules: psychological constructs adaptor modules (PCAM) and emotion adaptor modules (EAM), connected through a hierarchical bottom-up structure to ensure the dependency of emotion on psychological constructs.

PCAM. PCAMs consider the activation weights of the previous layer of BERT as the input. The activation weights are X_i^d where, (a) *d* represents the dimensionality of the hidden states transmitted between each layer of BERT, and (b) *i* represents the layer number. The encoder inside, essentially, augments the dimensionality of the input to *a* which helps retrieve latent features specific to the psychological context:

$$Enc(X_i^d) = X_i^d A^{d*a} + b^a \quad (2)$$

We use a BERT Multi-Head Attention mechanism once again as the feature extractor, but reduce the number of heads to $N/2$ as it was achieving competitive performance with fewer parameters. We therefore obtain $BA_{N/2}(Enc(X_i^d))$. Next, we apply a Decoder to resize the augmented data to the original hidden size *d* of BERT.

$$Dec(BA_{N/2}(Enc(X_i^d))) = BA_{N/2}(X_i^d)A^{a*d} + b^d \quad (3)$$

So, the output of PCAM is:

$$PCAM(X_i^d) = Dec(BA_{N/2}(Enc(X_i^d))) \quad (4)$$

EAM EAM is the adaptor module for classification on emotion, which generates outputs for PsyAM through one of three ways of feature fusion. First, pre-training fusion (*pre*) receives the last BERT layer's output and all the features from PCAMs. It defines them as a list of features: $[X_i^d, PCAM_1(X_i^d), PCAM_2(X_i^d)...])$. Next, it concatenates all the features, then passes it into Encoder inside EAM. So the feature propagation through EAM with fusion occurring *pre-training* looks like:

$$pre(X_i^d, PCAM_1(X_i^d), PCAM_2(X_i^d)...) = [X_i^d, PCAM_1(X_i^d), PCAM_2(X_i^d)...] \quad (5)$$

The output of residual from PsyAM is:

$$EAM(pre) = Dec(BA_{N/2}(Enc(pre(X_i^d, PCAM_1(X_i^d), PCAM_2(X_i^d)...)))) \quad (6)$$

Alternatively, when pre-training feature fusion is disabled, the post-training feature fusion methods (*post-linear* and *post-add*) receive the EAM's output and all the features from PCAMs, and then perform either an addition or a latent transformation of the concatenated features. The pooled output acts as the residual connection passed to the next BERT layer. In this case, the feature propagation through EAM looks like:

$$EAM(post-add) = Dec(BA_{N/2}(Enc(X_i^d))) + \sum_j PCAM_j(X_i^d) \quad (7)$$

$$EAM(post-linear) = LinearTransform(D) \quad (8)$$

$$D = post(Dec(BA_{N/2}(Enc(X_i^d))), PCAM_1(X_i^d), PCAM_2(X_i^d)...)) \quad (9)$$

B.1 Training Approach

After initializing the parameters for both BERT layers (pretrained BERT) and PsyAMs (Random), we train all of the psychological constructs tasks at the same time. Under the setting that (a) BERT layer parameters are frozen and (b) The EAM with feature fusion modules is disabled, these psychological constructs tasks contain their unique PCAM along with the BERT layer and classifier on the top, so they can be trained simultaneously without affecting the gradients of each other. Since feature fusion modules are disabled, inputs are directly propagated without transformation. Then the output of PCAM is passed to the output of BERT layer as residual:

$$Layer\_output = Rescale(FFN(BA_N(x)) + BA_N(x) + PCAM(X_i^d)) \quad (10)$$

After finishing training on psychological constructs and to realize a hierarchical bottom-up structure, we adjust the following settings: (a) Start the emotion training through the PCAMs to ensure the dependency of EAM on PCAM, (b) Choose a feature fusion strategy for the EAM, and (c) Freeze the parameters in all PCAMs. The output of EAM is passed to the output of BERT-PsyAM layer as the residual:

$$Layer\_output = Rescale(FFN(BA_N(x)) + BA_N(x) + EAM(X_i^d))) \quad (11)$$

For each task, the hidden states of the sequence input are transformed at each layer of BERT-PsyAM, but only the final hidden state of the *[CLS]* token is used for classification with the cross-entropy loss inside their unique classification head:

$$-\sum_{c=1}^{M} y_{o,c} \log(p_{o,c}) \quad (12)$$

where M is the number of classes, y is binary indicator (0 or 1), p is predicted probability observation o is of class c.

C New datasets' curation and annotation

HappyDB-2021 was collected through a Qualtrics Singapore panel sampled to obtain a gender- and age-representative split in terms of age and gender. Following the original data collection procedure, first we collected happy moments from survey respondents[12], together with their self reported duration of experiencing happiness. Participants were asked about three happy moments they experienced recently. Besides the happy moment, participants were also asked about the duration (i.e., the length) of happiness they experienced. The detailed instructions are sourced from (Asai et al., 2018) and are reported in Figure 5(a). Participants came from a representative distribution of gender (51.1% Female), age (M = 39.13; SD = 13.18), educational (Median = undergraduate degree), economic (Median = $7000 to $8999 monthly income), and racial backgrounds (83% majority). 984 happy moments thus collected were annotated through Amazon Mechanical Turk annotators, as reported in the next subsection.

TwitterUsers-2021 was collected from a Qualtrics USA panel sampled to obtain a nationally representative split in terms of age and gender. As a part of a larger survey, they were asked to share their Twitter handles and answer the Cantril Ladder question: "Think of a ladder, with the best possible life for you being a 10 and the worst possible life being a 0. Rate your own current life on that 0 to 10 scale." They also shared basic demographic information, such as their age, sex, level of education, and income. The detailed participant demographics are reported in Table 7.

Table 7: Demographic statistics of the TwitterUsers2021 dataset.

	TwitterUsers-2021
N	296
Gender	41.2% Female
Age group	
18-24 years	26
25-34 years	96
35-44 years	106
45-54 years	40
55-64 years	28
Highest education level	
No formal education	1
High school diploma	39
Some college	42
Technical Degree	32
Bachelor Degree	98
Graduate degree	68
PhD or equivalent	16
Annual household income	
less than $20,000	34
$20,000-$44,999	42
$45,000-$139,999	133
$140,000-$149,999	35
$150,000-$199,999	31
More than $200,000	21

C.1 Annotation

A sample of 1000 happy moments from HappyDB-2021 was published as an Amazon Mechanical Task to get five annotations per moment, to obtain annotations for agency and social interaction, following the same instructions as the original CLAff-HappyDB task. Annotators looked for evidence of:

- Personal agency: Describing whether or not the author was directly responsible for the happy moment that occurred. Example: "I made a nice birthday cake today."
- Social Interaction: Indicating whether or not other the happy moment involved other people. Example: "I had a good lunch with my mom."

The detailed annotator instructions are sourced

[12]The survey protocol for both surveys was approved by our university's Institutional Review Board.

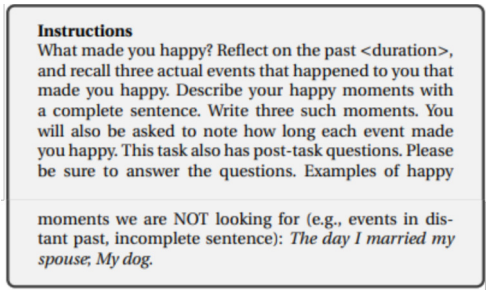
from (Jaidka et al., 2019) and are reported in Figure 5(b).

C.2 Semi-supervised labeling

We used a semi-supervised approach to annotate HappyDB-expand and the Kaggle Emotion datasets with Agency and Social Interaction labels with the help of the best-performing BERT-PsyAM classifiers trained on CLAff-HappyDB. By loading the pretrained parameters of Agency and Social task-specified psychological constructs adaptor modules(PCAMS), the BERT-PsyAM can easily predict psychological constructs labels. Furthermore, they can be applied to the main task of fine-tuning emotion classification in HappyDB-expand and the Kaggle Emotion datasets settings. Classification accuracies are reported in Table 8.

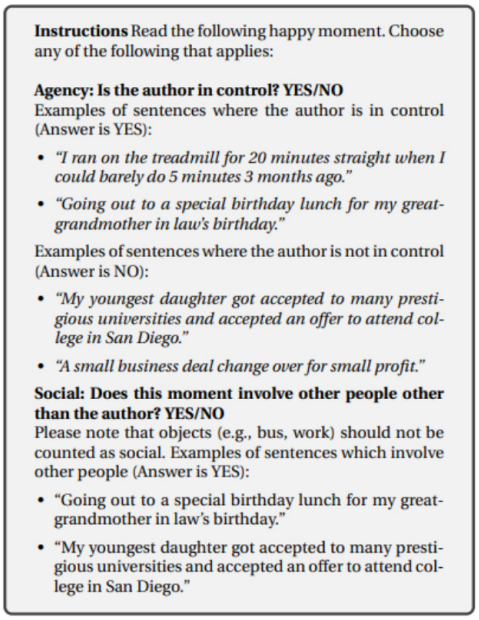
Table 8: Classification accuracies for models trained independently on agency and social interaction labels from CLAff-HappyDB. We used the BERT-PsyAM models to generate further labels on HappyDB-extended, SA-Emotions, and GoEmotions.

Approach	**Agency**	**Social**
Mlpclassifier	79.82	88.98
Bert base	84.89	91.54
Bert finetune	85.43	92.49
(Yu et al., 2021)	85.51	92.68
BERT-PsyAM	85.70	92.40



Instructions
What made you happy? Reflect on the past <duration>, and recall three actual events that happened to you that made you happy. Describe your happy moments with a complete sentence. Write three such moments. You will also be asked to note how long each event made you happy. This task also has post-task questions. Please be sure to answer the questions. Examples of happy moments we are NOT looking for (e.g., events in distant past, incomplete sentence): *The day I married my spouse; My dog.*

(a)



Instructions Read the following happy moment. Choose any of the following that applies:

Agency: Is the author in control? YES/NO
Examples of sentences where the author is in control (Answer is YES):

- *"I ran on the treadmill for 20 minutes straight when I could barely do 5 minutes 3 months ago."*
- *"Going out to a special birthday lunch for my great-grandmother in law's birthday."*

Examples of sentences where the author is not in control (Answer is NO):

- *"My youngest daughter got accepted to many prestigious universities and accepted an offer to attend college in San Diego."*
- *"A small business deal change over for small profit."*

Social: Does this moment involve other people other than the author? YES/NO
Please note that objects (e.g., bus, work) should not be counted as social. Examples of sentences which involve other people (Answer is YES):

- "Going out to a special birthday lunch for my great-grandmother in law's birthday."
- "My youngest daughter got accepted to many prestigious universities and accepted an offer to attend college in San Diego."

(b)

Figure 5: (a) Participant instructions to curate happy moments, sourced from the original HappyDB data collection (Asai et al., 2018).
(b) Annotation instructions to label happy moments with Agency and Social Interaction labels, sourced from the original CLAff-HappyDB task (Jaidka et al., 2019).

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Yes, in the Conclusion

☑ A2. Did you discuss any potential risks of your work?
Yes, in the Conclusion

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Yes, Section 4

☑ B1. Did you cite the creators of artifacts you used?
Yes, Section 4

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
We have included the link to the data repository where no PII is included.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Yes, in the Conclusion

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Yes, in the Appendix

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Yes, in Section 4 and in the Appendix

C ☑ Did you run computational experiments?

Yes, in Section 4 and 5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Yes, in Section 4 and 5

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Yes, in Section 4 and 5

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Yes, in Section 5

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Yes, in Section 4 and 5

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Yes in Section 4 and in the Appendix

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Yes in the Appendix

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Yes in Section 4 and in the Appendix

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Yes in the Appendix

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Yes in Section 4 and in the Appendix

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Yes in Section 4 and in the Appendix