A unified framework for cross-domain and cross-task learning of mental health conditions

Huikai Chua[♡] **Andrew Caines**[♠] **Helen Yannakoudakis**[♣♢]
[♡]Amazon Alexa
[♠]Department of Computer Science & Technology, University of Cambridge, U.K.
[♣]Department of Informatics, King's College London, U.K.
[♢]Kinhub
huikaic@amazon.co.uk andrew.caines@cl.cam.ac.uk helen.yannakoudakis@kcl.ac.uk

Abstract

The detection of mental ill-health based on an individual's use of language has received considerable attention in the NLP community. However, most work has focused on single-task and single-domain models, limiting the semantic space that they are able to cover and risking significant cross-domain loss. In this paper, we present two approaches towards a unified framework for cross-domain and cross-task learning for the detection of depression, post-traumatic stress disorder and suicide risk across different platforms that further utilizes inductive biases across tasks. Firstly, we develop a lightweight model using a general set of features that sets a new state of the art on several tasks while matching the performance of more complex task- and domain-specific systems on others. We also propose a multi-task approach and further extend our framework to explicitly capture the affective characteristics of someone's language, further consolidating transfer of inductive biases and of shared linguistic characteristics. Finally, we present a novel dynamically adaptive loss weighting approach that allows for more stable learning across imbalanced datasets and better neural generalization performance. Our results demonstrate the effectiveness of our unified framework for mental ill-health detection across a number of diverse English datasets.

1 Introduction

Depression is a mental health condition characterized by low mood, energy and self-esteem (American Psychiatric Association, 2013). One of the most serious effects of depression is the loss of joy in life, which leads to an increased suicide risk among people with depression.[1] However, due to the social stigma surrounding depression, many people who suffer from it hesitate to seek help. Suicide is one of the leading causes of death globally, especially among young people: it is the second most common cause of death among people aged 15–24.[2] Post-traumatic stress disorder (PTSD), which is characterized, among others, by symptoms of emotional outburst and negative thought, may also co-occur with depression, and can be a common response to PTSD sufferers.

There has thus been interest in the development of natural language processing (NLP) models for detection and/or prevention intervention. For example, this has been the focus of multiple shared tasks at the Computational Linguistics and Clinical Psychology (CLPsych) workshops (Coppersmith et al., 2015; Milne et al., 2016; Zirikly et al., 2019) as well as the Audio/Visual Emotion Challenge (AVEC) (Valstar et al., 2016; Ringeval et al., 2017, 2019).

However, previous work has tended to focus primarily on a single domain and/or mental health condition at a time. Each of the shared tasks listed above were focused on a single dataset from one domain; for example, the CLPsych 19 shared task used only forum posts from Reddit. The top systems at these shared tasks also frequently made use of domain-specific meta features such as the number of Reddit posts per time period, which were found to be among the most informative in suicide risk detection (Ruiz et al., 2019). Meanwhile, research has shown a lack of generalizability across datasets in classification models for mental health NLP (Harrigian et al., 2020).

The goal of our research is to develop models that can capture domain-independent and inter-related characteristics of different mental ill-health detection tasks, and generalize better. The novelty of our work is in proposing an alternative way of formulating the modeling of mental health conditions, which is more robust and effective compared to existing approaches, and we believe can benefit future research in this important task. We use En-

[1]https://www.who.int/en/news-room/fact-sheets/detail/suicide

[2]https://save.org/about-suicide/suicide-facts/

glish data from the CLPsych 2015 and 2019 shared tasks, which were obtained from Twitter and Reddit respectively, as well as the Distress Analysis Interview Corpus – Wizard of Oz (DAIC-WOZ) (Gratch et al., 2014), which consists of interview transcripts. Our open-domain setup precludes the use of domain-specific features such as the meta properties previously mentioned, as well as audio-visual cues from DAIC-WOZ interview recordings which may not always be available (e.g., due to user privacy concerns over voice and speech analysis).

To validate the general applicability of our approach, we experiment with two different types of approaches: 1) we develop novel multi-task learning architectures using a dynamically adaptive loss weighting scheduler that we show can lead to more effective learning across tasks/domains; and 2) we develop lightweight and interpretable models that, in contrast to the task-specific architecture and feature engineering used by many top shared task submissions (Mohammadi et al., 2019; Matero et al., 2019; Williamson et al., 2016), utilize a cross-task and cross-domain linguistic space that sets a new state of the art on several tasks while matching the performance of more complex task- and domain-specific systems on others.

To the best of our knowledge, this is the first approach towards a unified framework for open-domain (cross-domain) detection of different types of mental health conditions (cross-task).

2 Data & related work

We use data from two CLPsych shared tasks (Coppersmith et al., 2015; Zirikly et al., 2019), as well as the DAIC-WOZ corpus (Gratch et al., 2014) used in the AVEC challenges (Valstar et al., 2016; Ringeval et al., 2017, 2019), summarized below. While there has been little research in the development of cross-domain mental health models, there has been some effort to develop multi-task ones. These include models for different mental health conditions (anxiety, schizophrenia, panic, eating disorders) (Benton et al., 2017), or the use of auxiliary linguistic tasks such as figurative language detection (Yadav et al., 2020). However, all of the methods focus on a single domain (Twitter) and therefore capture a limited part of semantic space, whereas we focus on cross-domain methods that generalize across datasets.

2.1 CLPsych 15

The CLPsych 2015 shared task dataset (Coppersmith et al., 2015) was created by identifying Twitter users with depression or post-traumatic stress disorder (PTSD), based on whether they had publicly tweeted a diagnosis for either of these conditions. Each user is paired with an age- and gender-matched control, as estimated using the demographic classification tool from the World Well-Being Project (Sap et al., 2014). Up to the 3000 most recent tweets, excluding the original tweet of diagnosis, were collected for each user. The distribution is summarized in Table 4 in Appendix A.1.

The organizers set three binary classification tasks at the user level across each of three classes: **CD** (control vs depression), **CP** (control vs PTSD) and **DP** (depression vs PTSD). The best submission used supervised topic modelling and linear SVMs (Resnik et al., 2015).

2.2 CLPsych 19

The University of Maryland Reddit Suicidality Dataset (Version 2), used for CLPsych 19 (Zirikly et al., 2019; Shing et al., 2018), is made available with the assistance of the American Association of Suicidology. It contains the Reddit post history of 11,129 control users and another 11,129 users who have posted in r/SuicideWatch, a subreddit dedicated to supporting users who had or have suicidal thoughts. Of these users, 1097 were randomly sampled for annotation, with 993 annotated by crowdsourcing. These were then split into a training and test set as shown in Table 5 in Appendix A.1. Suicide risk has been annotated from 'None' to 'Severe' (given as character labels from 'a' to 'd').

The shared task organizers set three four-way classification tasks at the user level with different goals and restrictions on which posts may be used for classification: **Task A:** assessing a user's risk based on posts in the SuicideWatch reddit (typically a few posts per user); **Task B:** similar to Task A, but now all user posts, including those outside the SuicideWatch subreddit, may be used; **Task C:** this task is about screening users who may be at risk based on general posts (i.e., all posts *except* SuicideWatch posts may be used).

The best model on CLPsych 19 Tasks A and C used an SVM meta-classifier on top of eight sub-models based on CNN, RNN, bi-GRU and bi-

LSTM layers (Mohammadi et al., 2019). The sub-models utilized pretrained GloVe and ELMo word embeddings to produce user-level representations from user posts. An attention mechanism weighted each post based on their expected importance in predicting suicide risk. Finally, a fusion component weighted the user representations produced by each sub-model and then the SVM output the final predictions based on the weighted representations. The best model on CLPsych 19 Task B used various user-level post statistics (e.g., average unigram length, average unigrams per post) as well as information about the specific subreddits the users posted in, and processed posts separately based on that information. They also included a set of subreddit features, including one derived from popular subreddits, and one derived from subreddits distinctive of high-risk users.

2.3 DAIC-WOZ

The Distress Analysis Interview Corpus – Wizard of Oz (DAIC-WOZ) (Gratch et al., 2014) consists of transcribed interviews with veterans of the U.S. military and members of the general public from the Los Angeles area. The interviews were conducted using a virtual avatar controlled by a human interviewer, and then automatically transcribed with IBM Watson. The corpus includes Patient Health Questionnaire-8 (PHQ-8) scores for each participant as well as binary labels indicating depression. We utilize the binary labels and predict depression as a classification task. Although the corpus includes audio recordings and visual information such as facial and pose data, we opt not to make use of either audio or interviewer turns in DAIC-WOZ as we focus on modelling cross-domain user texts and task generalizability.

The best shared task classifier for DAIC-WOZ achieves 70% F1 (Williamson et al., 2016), in a submission to the AVEC 2016 challenge (Valstar et al., 2016). The approach used an ensemble model fusing predictions from audio, video, and semantic (text) features (e.g., task-specific audio features such as loudness variation and vocal tract physiology features). The authors also modeled the joint dynamical properties across facial action units using the video features, as well as included interviewer prompts in the text features, which they found to be more informative than user text alone.

2.4 GoEmotions

Previous work has found that fine-grained bag-of-emotions are useful features in depression detection in Reddit posts (Aragón et al., 2019). We extend the use of affective features across domains and different mental health conditions either in the form of lexicon-based emotion features (Section 3.1) or via the addition of an emotion detection auxiliary objective, **GoEmo** (Section 3.2), using the GoEmotions (Demszky et al., 2020) dataset. GoEmotions comprises around $58,000$ Reddit comments manually annotated using a fine-grained taxonomy of 27 emotions plus 'neutral', including a wide range of positive, negative and ambiguous emotions such as 'realization'. We use the released training, development and test splits,[3] consisting of 43410, 5426, and 5427 examples respectively. The number of examples per class ranges from over 5000 for the most frequent ('admiration'), to around 100 for the least frequent one ('grief').

2.5 Speaker characteristics

The datasets used are not necessarily balanced for representation. While the exact demographic labels are unavailable, CLPsych 15 is estimated to be roughly 80% white and nearly 90% female (Aguirre et al., 2021). Meanwhile, Reddit is estimated to be dominated by American users, which comprised nearly 50% of site traffic in 2020,[4] mostly male and under 25, according to a 2016 Reddit survey.[5] Therefore, it is likely that CLPsych 19 and GoEmotions, which are both collected from Reddit, follow similar demographic characteristics.

DAIC-WOZ features interviews with U.S. military veterans and residents of the Los Angeles area (Section 2) and thus was designed to specifically represent these social groups. However, in contrast to the CLPsych datasets, the gender distribution is approximately balanced between male and female (no other genders are declared in the dataset).

Overall, the corpora we use are dominated by young, male, North American users of social media. However, we note that annotator characteristics may differ. GoEmotions was annotated by native English speakers from India, while CLPsych 19 was annotated by crowdworkers from around the

[3] https://github.com/google-research/google-research/tree/master/goemotions/data

[4] https://www.statista.com/statistics/325144/reddit-global-active-user-distribution/

[5] https://www.reddit.com/r/dataisbeautiful/comments/5700sj/octhe_results_of_the_reddit_demographics_survey/

world on CrowdFlower.

3 Models

3.1 Lightweight feature-based model

Feature-based models have been shown to achieve state of the art results on various mental ill-health detection tasks, while can facilitate model interpretability, which is crucial in high-stakes areas such as mental health. We focus on the development of such a lightweight approach that furthermore captures shared and generalizable properties across tasks and domains. In contrast to the task-specific architecture and feature engineering used by many top shared task submissions (Mohammadi et al., 2019; Matero et al., 2019; Williamson et al., 2016), we utilize the datasets' development sets to identify the most effective set of domain-invariant features.

Our best model uses tf–idf word unigrams, character (2,4)-grams, and part-of-speech (POS) tags based on NLTK's POS tagger (Bird and Loper, 2004). Inspired by previous work, we also add the following count-based features: first-person singular pronouns which have been identified as more frequently used among depressives (Al-Mosaiwi and Johnstone, 2018) across demographic lines (Edwards and Holtzman, 2017) due to increased self-focus (Wolohan et al., 2018; Brockmeyer et al., 2015); first-person plural pronouns (although depressed people might use the first-person singular more often, they might not necessarily express as much social engagement; De Choudhury et al. (2013)); words reflecting absolutist thinking (Al-Mosaiwi and Johnstone, 2018) such as 'always', as cognitive rigidity has been linked to suicidal ideation (Ellis and Rutherford, 2008).

We also calculate sentence-level sentiment scores using NLTK's VADER tool (Hutto and Gilbert, 2015) and include the average over all sentences; as well as emotion features based on the NRC Word–Emotion Association Lexicon (EmoLex) (Mohammad and Turney, 2013; Mohammad, 2011; Mohammad and Yang, 2011; Mohammad and Turney, 2010). EmoLex comprises eight emotions – anger, anticipation, disgust, fear, joy, sadness, surprise, trust – as well as negativity and positivity sentiment dimensions. To identify the most predictive affective characteristics of text among those 10 features, we perform grid search on the development data, find anger, joy, surprise, positivity and negativity to be the most discriminative, and include these in our final model.[6]

We experiment with two lightweight models, support vector machines (SVMs) (Cortes and Vapnik, 1995) and gradient-boosted decision trees (GBDTs) from XGBoost (Chen and Guestrin, 2016). During tuning, we find the latter to give the best performance (between 3-20 F1 points difference) and we therefore choose this for our experiments.

3.2 MT-DNN model

We develop a multi-task deep neural network (MT-DNN) (Liu et al., 2019c,b; He et al., 2019; Liu et al., 2019a; Jiang et al., 2020; Liu et al., 2020; Wang et al., 2019; Cheng et al., 2020) [7] to directly leverage inductive transfer between our tasks. Our model consists of a pre-trained shared encoder followed by separate task-specific layers. We use the uncased English BERT$_{BASE}$ model provided by Hugging Face (Devlin et al., 2019; Wolf et al., 2020) as the encoder shared across the different datasets, encode the most recent 512 tokens[8] and use the [CLS] token as the post embedding for classification. The task-specific layers consist of a linear layer with either a sigmoid or softmax activation for the multi-label (GoEmotions) and classification tasks respectively, and the model is optimized using (binary) cross entropy. The shared encoder makes up the bulk of the MT-DNN model, with around 110 million parameters.

The various datasets differ greatly in the number of examples per class. We find that running the model for 30 epochs ensures that all have had a chance to converge.[9] This is further discussed in Section 5. To improve stability, we accumulate gradients over 3 steps during training, using a batch size of 16. We manually tune the learning rate to 9e-5 on the development set using the Adamax optimizer (Kingma and Ba, 2017).

Adaptive loss weights The different datasets have different distributions and learning curves, making it difficult to determine an appropriate stop-

[6] The use of the two sentiment scores improved performance further to the averaged VADER scores; we surmise this is due to the more fine-grained information added via the explicit counts of positivity and negativity expressed in a post.

[7] https://github.com/namisan/mt-dnn

[8] Word boundaries were respected, i.e., if the n most recent words have a subtoken length greater than 512, then only the $(n-1)$ most recent words were used.

[9] 30 epochs takes around 2 hours to train the multi-task model on a Tesla P100 on CLPsych 15, CLPsych 19 and DAIC-WOZ. Adding GoEmotions (substantially more examples than any of the other tasks) increases runtime to around 4 hours.

ping criterion for the multi-task model. While we can train the model until the slowest task has peaked on the development data (as mentioned above), this is likely to lead in overfitting for the other tasks. On the other hand, sequential training of tasks runs the risk of catastrophic forgetting. To mitigate this, we weight the losses of each task in the multi-task model, and propose an approach that dynamically adapts the weights based on whether a task has reached convergence or not.

We start by initializing the weights to 1, and set patience P to 3 epochs if the development F1 improves at the end of each epoch. If a task stops improving but has already reached 90% of a predetermined target performance T, its loss weight is gradually reduced by 0.1 with a lower bound of 0.5 to ensure it is always weighted at least as much as the auxiliary emotion task (see below). On the other hand, if a task has not improved over P epochs and has yet to reach 50% of the target performance, its loss weight is multiplied proportional to $\frac{T}{S}$, where S refers to the current performance, up to a maximum of 1.5 times.

We experiment with two different ways of setting the target performance that a model should reach before its loss weight is adjusted: **Adapt-Fixed** that uses a fixed target performance threshold of 80% F1 for each task (we manually tuned this on the development set and found this to perform best); **Adapt-Variant** where the target for each task is set separately. Here, we first complete an initial MT-DNN run with all tasks and constant weights. The individual task target performance threshold is then the best development F1 achieved for that task at any epoch. To assess the effectiveness of our weighting approaches, we furthermore report results without adaptive loss weights, referred to as **Constant**, but where all tasks are equally weighted and each weight is set to 1. The only exception to the above is the auxiliary emotion detection task, GoEmo, for which the loss weight w is downweighted and fixed at 0.5 to prioritize performance on the mental health tasks. We found that the model converged to roughly the same F1 score on GoEmo as when $w = 1.0$, but resulted in better performance on our main tasks.

While existing multi-task approaches may adapt the scheduler such that texts from underperforming tasks are selected more often (Jean et al., 2019), sampling tasks effectively presents a challenge in our setting, characterized by high variation in class distribution and dataset sizes. The latter range from thousands of examples per class (GoEmotions) to less than a hundred (DAIC-WOZ). Our approach presents a simpler alternative to ameliorating this problem that does not rely on explicit data manipulation but rather directly exploits the learning patterns of a given model.[10]

3.3 Single-task baselines

We include single-task BERT-based baselines trained on each of the datasets separately, using a linear schedule with 20 warmup steps, a learning rate of 5e-5, a batch size of 16, and no gradient accumulation. The models are trained until F1 does not improve on the development data over 3 consecutive epochs and the best model is selected.

4 Results

Experimental setting For DAIC-WOZ and GoEmotions, we use the published train/dev/test splits. Since CLPsych 15 and 19 did not include a development set, we put aside 10% of the training sets (randomly selected and stratified by class) for development. Both approaches use the same data and are tuned across the task development sets. We evaluate model performance using macro F1. For CLPsych 15, we also report precision separately, as this was the *primary* evaluation metric of the shared task. We also report F1*, the *official* metric for CLPsych 19 Task C, which is F1 computed without the lowest risk class ('a').

Feature-based model In Table 1, we can see that the Feature Model achieves, overall, a high performance across tasks using its set of domain-invariant features for predicting the different types of mental health conditions. We report macro-F1 for all tasks, along with precision for the CLPsych 15 tasks (the *official* evaluation metric used), and F1* (without class 'a') for CLPsych 19 task C (the task's *official* evaluation metric). F1 was not reported for the CLPsych 15 shared tasks, but we have rather estimated it (†) based on the reported precision and ROC curve. However, we can calculate the average macro-F1 performance across tasks for our models (last column; based on the F1 value for Task C). We can see that our performance comes close to or surpasses the best shared task results on all tasks

[10] While other approaches to scheduling can be investigated (e.g., Kiperwasser and Ballesteros (2018)), our main aim is to demonstrate the validity and robustness of a unified approach, and therefore leave this for future work.

	CD		CP		DP		A	B	C		DW	Avg
Model	F1	Prec	F1	Prec	F1	Prec	F1	F1	F1	F1*	F1	F1
Shared Task	**84**†	86	87†	89	69†	83	48.1	**47.0**	–	**26.8**	**70.0**	–
Feature Model	80.2	**87.8**	**92.1**	**94.9**	**83.7**	**86.3**	47.9	33.4	31.2	24.0	68.1	**62.4**
Single Task	52.8	62.3	69.7	67.8	61	71.1	41.1	25.2	33.3	19.0	41.9	46.4
epoch		(9)		(5)		(4)	(7)	(2)		(3)	(13)	
Constant	57.1	51.9	62.7	62.7	56.1	68.9	46.3	23.7	29.1	13.7	46.8	45.9
Adapt-Fixed	57.5	53.1	59.7	64.8	57.1	77.3	47.5	27.5	33.0	18.2	58.1	48.6
Adapt-Variant	56.1	50.0	61.6	63.4	53.9	71.4	**50.5**	27.3	32.3	17.1	40.0	46.0
Constant$_{GoEmo}$	58.3	52.7	64.6	68.2	58.1	69.4	38.8	30.8	**35.9**	22.1	47.0	47.6
Adapt-Fixed$_{GoEmo}$	57.9	53.4	62.7	62.7	58.7	72.5	49.3	27.7	33.5	18.7	60.5	53.6
Adapt-Variant$_{GoEmo}$	57.0	53.2	62.6	63.8	58.0	74.7	48.6	27.6	33.0	18.2	63.2	50.0

Table 1: Model performance across datasets: CD, CP and DP from CLPysch 15; Tasks A, B and C from CLPsych 19; and DW, the binary classification task on DAIC-WOZ. We report macro-F1 for all tasks, along with precision for the CLPsych 15 tasks (the *official* evaluation metric used), and F1* (without class 'a') for CLPsych 19 task C (the task's *official* evaluation metric). F1 was not reported for the CLPsych 15 shared tasks, but we have estimated it (†) based on the reported precision and ROC curve. However, we show the average macro-F1 performance across tasks for our models (last column; based on the F1 value for Task C). *Shared Task* represents the current state-of-the-art performance. 'Feature Model' is our feature-based baseline, while 'Single Task' is BERT fine-tuned to each task individually, also showing the epoch at which training was halted according to our early stopping criterion (see Section 5). 'Constant' is the multi-task model without adaptive loss weights; Adapt-Fixed and Adapt-Variant refer to the different versions of our adaptive loss weighting algorithm. 'GoEmo' indicates the addition of the emotion objective using the GoEmotions dataset. The best performance in each column is highlighted in bold.

except Task B, where we did not make use of additional contextual information about the subreddit the post belongs to (Matero et al., 2019).[11]

Single-task baselines We can see that the single-task BERT baselines failed to outdo the best shared task scores on all tasks, performing especially poorly on CLPsych 15 (CD, CP, DP).[12]

MT-DNN The baseline multi-task model, Constant (without use of adaptive loss weighting), showed mixed results with F1 improvements in CD, A and DW, but decreases compared to the single-task model on the other tasks. This confirms the need for a unified (neural) approach that directly takes into account the training dataset distributions and learning curves. Our adaptive loss weight algorithms, Adapt-Fixed and Adapt-Variant, attempt to ameliorate this. Specifically, we find that Adapt-Fixed can balance the performance across multiple tasks and achieve better overall performance (avg F1) compared to both the single-task and Constant counterparts, contributing to an effective unified approach. On the other hand, Adapt-Variant is on par with Constant. The effectiveness of Adapt-Fixed can be attributed to the fact that it enforces learning to a certain (high) level of performance for each task, as opposed to Adapt-Variant that has a less strict approach to learning performance thresholds.

Adding emotion detection (GoEmo) as an auxiliary task leads to overall improvements (avg F1). Comparing the Constant variants with and without GoEmo, we see the largest improvements in Tasks B and C of 7.1 and 8.4 F1 points respectively. This can be explained by the fact that both the GoEmotions and CLPsych 19 datasets were collected from Reddit (however, it seems that the dataset generalizes more poorly to Task A, which was collected only from one specific subreddit). Overall, we observe again that Adapt-Fixed achieves the best performance across the neural models, with particularly large improvements in Task A as well as DW of 10.5 and 13.5 points respectively.

5 Discussion

MT-DNN vs. Feature Model The feature-based model showed the best performance across all tasks utilizing its set of domain-invariant features, demonstrating that they share a common linguis-

[11] Since subreddits are typically organized around common interests or shared experiences, these provide valuable contextual information about a person's background. For example, we might deduce that someone who frequents the 'ukpolitics' subreddit most likely lives in the UK and is interested in politics. Such information however is not always readily available in other social media such as Twitter, where tweets are posted on users' walls instead of being organized into sub-forums.

[12] To test whether this can be attributed to the BERT text length restriction, we experimented with additional models such as longformers (Beltagy et al., 2020); however, BERT was nevertheless found to perform best.

Model	CD F1	CP F1	DP F1	A F1	B F1	C F1	DW F1	Avg F1
Constant	60.5	65	70.5	49.5	29.9	33.6	53.8	48.1
epoch	(10)	(13)	(5)	(17)	(16)	(16)	(20)	(12)
Adapt-Fixed	57.5	59.7	57.1	47.5	27.5	33	58.1	48.6
Constant$_{GoEmo}$	58.5	65.7	70.7	51.6	32.9	36.8	59.3	49.5
epoch	(9)	(4)	(5)	(5)	(10)	(4)	(24)	(24)
Adapt-Fixed$_{GoEmo}$	57.9	62.7	58.7	49.3	27.7	33.5	60.5	53.6

Table 2: The highest F1 attained by Constant and Constant$_{GoEmo}$ for each of the tasks separately, together with the epoch at which this is observed. The epoch with the best average F1 score is included under 'Avg'. For comparison, the Adapt-Fixed (with and without GoEmo) results at epoch 30 are reproduced from Table 1.

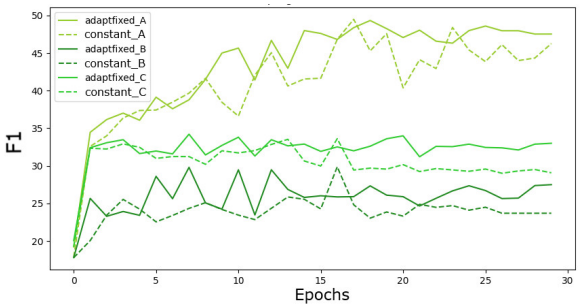

Figure 1: F1 score during training for 30 epochs for each task in CLPsych 19 for the Constant (dotted) and Adapt-Fixed (solid) models. The graphs for all datasets are reproduced in the Appendix A.4.

tic/feature space.[13] Therefore, using a multi-task model should theoretically enhance performance by allowing the shared encoder to simultaneously learn features at different levels of abstraction from all tasks at once. However, the Constant model achieved a lower average F1 score than the single-task BERT variants. This can be attributed to the difficulty of balancing the performance of multiple different tasks, where each have different learning schedules. While Adapt-Fixed provides a solution to this, it seems that, overall, there is scope for improvement on bridging the gap between feature-based and neural approaches for this task.[14] To better understand the effect of Adapt-Fixed on neural performance, we perform a detailed analysis below, illustrating the challenges in balancing different tasks and datasets within the neural approach.

Adaptive loss weighting analysis The adaptive loss weighting algorithm is motivated by the very different learning schedules of the different tasks. As can be seen in Table 1 (row *epoch*), the single task models were all stopped at different epochs for the different tasks, ranging from 2 epochs for B to 13 for DW. The difference in learning schedules is amplified in the multi-objective MT-DNN model. In Table 2, we report the highest F1 attained by the Constant and Constant$_{GoEmo}$ models for each of the tasks separately, together with the epoch at which this is observed. We can see that the best individual task F1 occurs at different epochs but with a wider spread. These range from 5 to 20 and from 4 to 24 for the Constant and Constant$_{GoEmo}$ models respectively. Therefore, in order to ensure the model is able to learn all tasks, we train it for a total of 30 epochs (compared to around 5 typically used for BERT (Devlin et al., 2019)), additionally utilizing the adaptive loss weighting algorithm to reduce overfitting.

In Section 4, we noted that the Adapt-Fixed models generally improved both single task as well as average F1 after 30 epochs of training. In Table 2, we can also see they outperform the highest F1 average attained at *any* epoch by the Constant models (for ease of comparison, we include the Adapt-Fixed and Adapt-Fixed$_{GoEmo}$ test results at epoch 30, reproduced from Table 1). Comparing individual task F1s, both the Adapt-Fixed versions scored within 3 points of the best achieved Constant F1 for 5 out of 7 tasks. This shows that the algorithm has been successful in balancing performance across most of the tasks, while improving the overall F1.

Qualitatively, we also observe that the adaptive loss weighting algorithm has a smoothing effect on training (Figure 1). Comparing the Constant model (dotted lines) to Adapt-Fixed (solid lines), we can see that, using the latter, we obtain a more stable version which empirically converges faster.

[13] Running a set of ablation studies, we find tf–idf unigrams, char ngrams and POS ngrams to be highly predictive.

[14] Notably, the neural model outperformed the feature-based model on Task A, where it also outperformed the state-of-the-art. As the most specialized task, consisting only of posts related to mental health, it is likely that Task A benefitted the most from information learned from the other related tasks.

Model	CD F1	CP F1	DP F1	A F1	B F1	C F1	DW F1	Avg F1
Positive	56.9	62.2	60.5	45.2	27.7	34.4	8.7	42.2
Negative	57.8	64.1	60.3	41.2	30.1	33.9	51.6	48.4
GoEmo	58.3	64.6	58.1	38.8	30.8	35.9	47.0	47.6

Table 3: F1 of the Constant$_{GoEmo}$ MT-DNN model using only positive and only negative emotions. For comparison, its performance with all emotion classes (GoEmo) is reproduced from Table 1.

Emotion features The GoEmotions auxiliary objective seems particularly beneficial, resulting in improvements over the single-task models and across all MT-DNN variants, with the highest F1 observed using the Adapt-Fixed model, leading to an average increase of 5% across datasets compared to its no-emotion counterpart. The feature-based model captures affective characteristics of language explicitly via the use of EmoLex features, as well as implicitly via the use of word and character ngrams. In qualitative analyses we find that, among the most highly predictive features, there exist affective terms such as 'pissed', 'bloody' and 'endure' for Task A (moderate and severe suicide risk) and 'loves' (low suicide risk); and 'afraid' and 'annoying' for DAIC-WOZ (depressed class).

To investigate the effect that negative emotions specifically might have in the detection of mental ill-health, we separate the 28 emotion labels into positive (13 total) and negative classes (11 total)[15] and now use these to train Constant$_{GoEmo}$. In Table 3, we can see that, overall, the exclusive use of negative emotions leads to an increased Avg F1 across all datasets. Notably, the effect is substantial for DW, with a 32.9 point difference compared to using positive emotions. In Appendix A.3, we also investigate learning effects in the opposite direction and examine how mental ill-health detection might affect performance of emotion detection.

6 Conclusion

We presented two approaches to cross-domain and cross-task mental ill-health detection. The first involves the development of a general set of features; the second uses a multi-task model, utilizing BERT as a shared encoder (Devlin et al., 2019). We found the former to perform well across all domains and tasks, demonstrating that they share a common set of linguistic cues. In comparison to shared task submissions which use complex neural models (Mohammadi et al., 2019; Matero et al., 2019; Williamson et al., 2016), our approach either matches their performance or improves over state-of-the-art results using a lightweight decision tree-based model. Such models are furthermore more transparent and interpretable with respect to the basis upon which they make predictions, which is crucial in high-stakes domains such as mental health and in assessing model validity and whether it measures what is intended to be measured.

We furthermore investigated the use of affective features, as well as examined negative emotion features in isolation, as a useful inductive bias for the detection of different types of mental health conditions, extending previous work that examines the effect of emotion features in a single domain and single task setting (depression detection in Reddit posts; Aragón et al. (2019)). Emotion detection, as an auxiliary objective, increased the average F1 score by 1.7 points, with the most substantial improvements observed in tasks from the same domain as the emotion dataset (CLPsych 19).

Finally, we presented an adaptive loss weighting algorithm which successfully balances performance across tasks with different learning schedules while increasing the overall performance. A comparison of model results with and without adaptive weighting revealed that it not only led to improved performance, but also outperformed the best average F1 score achieved over all epochs with constant weighting.

Our feature-based approach outperformed the neural counterpart, indicating that there is scope for further research towards a unified framework for open-domain detection of various mental health conditions. However, our feature-based results experimentally demonstrate that such an approach is feasible and effective, and achieves a new state of the art on several tasks.

To the best of our knowledge, this is the first approach towards reformulating mental ill-health detection within a unified framework. Our paper aims to lay a platform for future research, facilitating progress in this important effort.

[15] Four classes – confusion, realization, pride, and neutral – are excluded as they are not overtly positive or negative.

7 Ethical concerns

This project was reviewed and approved by the Department of Computer Science & Technology Ethics Committee, University of Cambridge.

Risks that may arise from this work include perpetuation of biases existing in the datasets used. Gender labels for each participant are unavailable in all except the DAIC-WOZ dataset, so the distribution may not be balanced. Crucially, mostly American speakers and users were included in the creation of the datasets. As cross-cultural differences have been found in the way people express depression (Loveys et al., 2018), further work would be required to investigate whether the approaches adopted generalize to datasets across demographic lines.

Such concerns also arise from the use of large language models, as it may be more difficult to correct bias in the large amounts of language data used for training (Blodgett et al., 2020). It has also been shown that it is possible to recover the original training texts from large language models (Carlini et al., 2020). Therefore, deployment of any system including such language models, such as the multi-task variants presented herein, should ensure not to compromise the privacy of the user. However, we note that the datasets used here have all been anonymized.

Finally, developers of models that can flag users should also consider the purpose of such predictions as well as whether they can be used to take actions against users; e.g., as part of 'social media checks' when screening job applicants. While well-intending friends and family members might use them to help those anxious about seeking help, others might also use such tools to discredit or slander others, particularly in cultures where mental health conditions are still stigmatized.

References

Carlos Aguirre, Keith Harrigian, and Mark Dredze. 2021. Gender and racial fairness in depression research using social media. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2932–2949, Online. Association for Computational Linguistics.

Mohammed Al-Mosaiwi and Tom Johnstone. 2018. In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. *Clinical Psychological Science*, 6(4):529–542. PMID: 30886766.

American Psychiatric Association. 2013. *Diagnostic and statistical manual of mental disorders: DSM-5*, 5th ed. edition. American Psychiatric Association, Arlington, VA.

Mario Ezra Aragón, Adrian Pastor López-Monroy, Luis Carlos González-Gurrola, and Manuel Montes-y Gómez. 2019. Detecting depression in social media using fine-grained emotions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1481–1486, Minneapolis, Minnesota. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017. Multitask learning for mental health conditions with limited social media data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 152–162, Valencia, Spain. Association for Computational Linguistics.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Timo Brockmeyer, Johannes Zimmermann, Dominika Kulessa, Martin Hautzinger, Hinrich Bents, Hans-Christoph Friederich, Wolfgang Herzog, and Matthias Backenstrass. 2015. Me, myself, and I: self-referent word use as an indicator of self-focused attention in relation to depression and anxiety. *Frontiers in Psychology*, 6:1564.

Nicholas Carlini, Florian Tramèr, Eric Wallace, M. Jagielski, Ariel Herbert-Voss, K. Lee, Adam Roberts, Tom Brown, D. Song, Ú. Erlingsson, Alina Oprea, and Colin Raffel. 2020. Extracting training data from large language models. *ArXiv*, abs/2012.07805.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. *CoRR*, abs/1603.02754.

Hao Cheng, Xiaodong Liu, Lis Pereira, Yaoliang Yu, and Jianfeng Gao. 2020. Posterior differential regularization with f-divergence for improving model robustness. *arXiv preprint arXiv:2010.12638*.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. CLPsych 2015 shared task: Depression and PTSD on Twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39, Denver, Colorado. Association for Computational Linguistics.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. AAAI.

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, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

To'Meisha Edwards and Nicholas S. Holtzman. 2017. A meta-analysis of correlations between depression and first person singular pronoun use. *Journal of Research in Personality*, 68:63–68.

Thomas Ellis and Billy Rutherford. 2008. Cognition and suicide: Two decades of progress. *International Journal of Cognitive Therapy*, 1:47–68.

Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, and Louis-Philippe Morency. 2014. The Distress Analysis Interview Corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2020. Do models of mental health based on social media data generalize? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3774–3788, Online. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Weizhu Chen, and Jianfeng Gao. 2019. A hybrid neural network model for commonsense reasoning. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 13–21, Hong Kong, China. Association for Computational Linguistics.

C.J. Hutto and Eric Gilbert. 2015. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*.

Sébastien Jean, Orhan Firat, and Melvin Johnson. 2019. Adaptive scheduling for multi-task learning. *CoRR*, abs/1909.06434.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. 2020. SMART: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2177–2190, Online. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Eliyahu Kiperwasser and Miguel Ballesteros. 2018. Scheduled multi-task learning: From syntax to translation. *CoRR*, abs/1804.08915.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. 2019a. On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019b. Improving multi-task deep neural networks via knowledge distillation for natural language understanding. *arXiv preprint arXiv:1904.09482*.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019c. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.

Xiaodong Liu, Yu Wang, Jianshu Ji, Hao Cheng, Xueyun Zhu, Emmanuel Awa, Pengcheng He, Weizhu Chen, Hoifung Poon, Guihong Cao, and Jianfeng Gao. 2020. The Microsoft toolkit of multi-task deep neural networks for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 118–126, Online. Association for Computational Linguistics.

Kate Loveys, Jonathan Torrez, Alex Fine, Glen Moriarty, and Glen Coppersmith. 2018. Cross-cultural differences in language markers of depression online. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 78–87, New Orleans, LA. Association for Computational Linguistics.

Matthew Matero, Akash Idnani, Youngseo Son, Salvatore Giorgi, Huy Vu, Mohammad Zamani, Parth Limbachiya, Sharath Chandra Guntuku, and H. Andrew Schwartz. 2019. Suicide risk assessment with multi-level dual-context language and BERT. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 39–44, Minneapolis, Minnesota. Association for Computational Linguistics.

David N. Milne, Glen Pink, Ben Hachey, and Rafael A. Calvo. 2016. CLPsych 2016 shared task: Triaging content in online peer-support forums. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 118–127, San Diego, CA, USA. Association for Computational Linguistics.

Saif Mohammad. 2011. From once upon a time to happily ever after: Tracking emotions in novels and fairy tales. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 105–114, Portland, OR, USA. Association for Computational Linguistics.

Saif Mohammad and Peter Turney. 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34, Los Angeles, CA. Association for Computational Linguistics.

Saif Mohammad and Peter Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29.

Saif Mohammad and Tony Yang. 2011. Tracking sentiment in mail: How genders differ on emotional axes. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*, pages 70–79, Portland, Oregon. Association for Computational Linguistics.

Elham Mohammadi, Hessam Amini, and Leila Kosseim. 2019. CLaC at CLPsych 2019: Fusion of neural features and predicted class probabilities for suicide risk assessment based on online posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 34–38, Minneapolis, Minnesota. Association for Computational Linguistics.

Philip Resnik, William Armstrong, Leonardo Claudino, and Thang Nguyen. 2015. The University of Maryland CLPsych 2015 shared task system. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 54–60, Denver, Colorado. Association for Computational Linguistics.

Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, Siyang Song, Shuo Liu, Ziping Zhao, Adria Mallol-Ragolta, Zhao Ren, Mohammad Soleymani, and Maja Pantic. 2019. Avec 2019 workshop and challenge: State-of-mind, detecting depression with ai, and cross-cultural affect recognition. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, AVEC '19, page 3–12, New York, NY, USA. Association for Computing Machinery.

Fabien Ringeval, Björn Schuller, Michel Valstar, Jonathan Gratch, Roddy Cowie, Stefan Scherer, Sharon Mozgai, Nicholas Cummins, Maximilian Schmitt, and Maja Pantic. 2017. AVEC 2017: Real-life depression, and affect recognition workshop and challenge. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, AVEC '17, pages 3–9, New York, NY, USA. ACM.

Victor Ruiz, Lingyun Shi, Wei Quan, Neal Ryan, Candice Biernesser, David Brent, and Rich Tsui. 2019. CLPsych2019 shared task: Predicting suicide risk level from Reddit posts on multiple forums. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 162–166, Minneapolis, Minnesota. Association for Computational Linguistics.

Maarten Sap, Gregory Park, Johannes Eichstaedt, Margaret Kern, David Stillwell, Michal Kosinski, Lyle Ungar, and Hansen Andrew Schwartz. 2014. Developing age and gender predictive lexica over social media. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1146–1151, Doha, Qatar. Association for Computational Linguistics.

Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. 2018. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 25–36, New Orleans, LA. Association for Computational Linguistics.

Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. 2016. AVEC 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, AVEC '16, pages 3–10, New York, NY, USA. ACM.

Dilin Wang, Chengyue Gong, and Qiang Liu. 2019. Improving neural language modeling via adversarial training. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6555–6565. PMLR.

James R. Williamson, Elizabeth Godoy, Miriam Cha, Adrianne Schwarzentruber, Pooya Khorrami, Youngjune Gwon, Hsiang-Tsung Kung, Charlie Dagli, and Thomas F. Quatieri. 2016. Detecting

depression using vocal, facial and semantic communication cues. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, AVEC '16, page 11–18, New York, NY, USA. Association for Computing Machinery.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online. Association for Computational Linguistics.

JT Wolohan, Misato Hiraga, Atreyee Mukherjee, Zeeshan Ali Sayyed, and Matthew Millard. 2018. Detecting linguistic traces of depression in topic-restricted text: Attending to self-stigmatized depression with NLP. In *Proceedings of the First International Workshop on Language Cognition and Computational Models*, pages 11–21, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Shweta Yadav, Jainish Chauhan, Joy Prakash Sain, Krishnaprasad Thirunarayan, Amit Sheth, and Jeremiah Schumm. 2020. Identifying depressive symptoms from tweets: Figurative language enabled multitask learning framework. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 696–709, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. 2019. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 24–33, Minneapolis, Minnesota. Association for Computational Linguistics.

A Appendix

In the appendix, we have some supplementary statistics about the datasets and training performance. The first three tables show the label distributions for each of the datasets. We also include the emotions from GoEmotions with the largest change in performance between a single task GoEmotions model and the full GoEmo Constant model. Finally, we include graphs comparing the F1 score progress between the Constant and Adapt-Fixed training algorithms per dataset.

Label	Training size	Test size
Control	572	300
Depression	327	150
PTSD	246	150

Table 4: CLPsych 15 dataset statistics for the number of users per class and data split.

Label	Training size	Test size
a (None)	127	36
b (Low)	50	50
c (Moderate)	113	115
d (Severe)	206	44
Control	497	124

Table 5: CLPsych 19 dataset statistics for the number of users per class and data split.

Label	Training	Dev	Test
0 (Non-depressed)	76	23	33
1 (Depressed)	30	12	14

Table 6: DAIC-WOZ dataset statistics for the number of participants per class and data split.

A.1 Dataset statistics

A.2 Computing infrastructure and run-time

The full set of features for the feature-based model can be (pre)computed within hours on CPU. Training for the SVMs and GBDTs usually completes within an hour. The MT-DNN model is essentially the size of the shared encoder, with around 110 million parameters). 30 epochs takes around 2 hours on GPU (Tesla P100) for the core set of CLPsych 15, CLPsych 19 and DAIC-WOZ tasks. Adding GoEmotions, which contains substantially more examples than any of the core tasks, increases runtime to around 4 hours.

A.3 Effect of mental health datasets on emotion detection

To investigate the effect the learning of mental health conditions might have on emotion detection performance, we also train a single-task BERT baseline on the GoEmotions dataset and compare the F1 scores for each of the emotion classes to Constant$_{GoEmo}$. Only 3 out of the 28 emotion classes see a decrease in performance between the single-task BERT and MT-DNN model: neutral (-1.6), realization (-2.7), and pride (-10.0). The emotions which have F1 changes of 10 or more points are presented in Table 7. As can be seen, most of the emotions with substantial F1 improvements are positive emotions. This is rather surprising, as mental ill-health such as depression is typ-

Emotion	Single Task	MT-DNN	Change
Nervousness	8.0	31.6	23.6
Desire	29.9	47.8	17.9
Caring	25.7	43.4	17.7
Relief	0.0	15.4	15.4
Joy	50.6	62.9	12.3
Disappointment	19.5	31.6	12.1
Approval	27.6	39.4	11.8
Pride	40.0	30.0	-10.0
Avg	41.2	47.1	5.9

Table 7: Emotion detection performance (F1) for classes that are affected the most with and without multi-task learning (MT-DNN Constant$_{GoEmo}$ and Single Task emotion detection respectively). The bottom row presents the average F1 score over *all* 28 emotion classes.

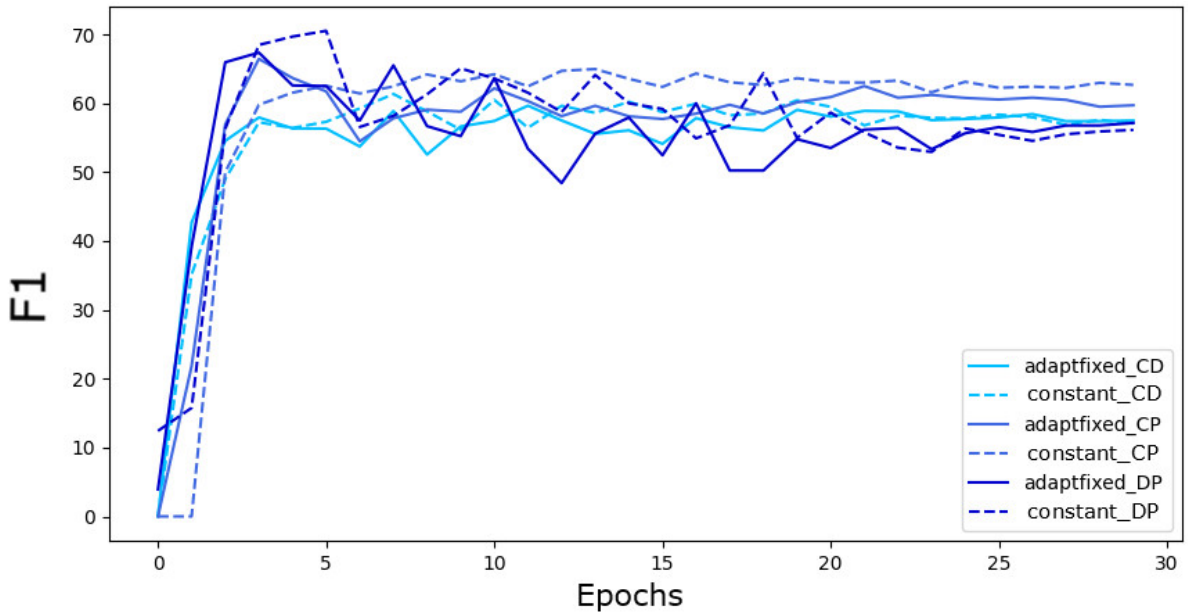

Figure 2: F1 score during training for 30 epochs for each task in CLPsych 15 for the Constant (dotted) and Adapt-Fixed (solid) models.

ically associated with negative emotions (Aragón et al., 2019); however, the datasets aggregate information across control groups too, which can present useful additional features for the detection of positive emotions and the absence of mental ill-health.

A.4 Graphs of F1 score progress during training

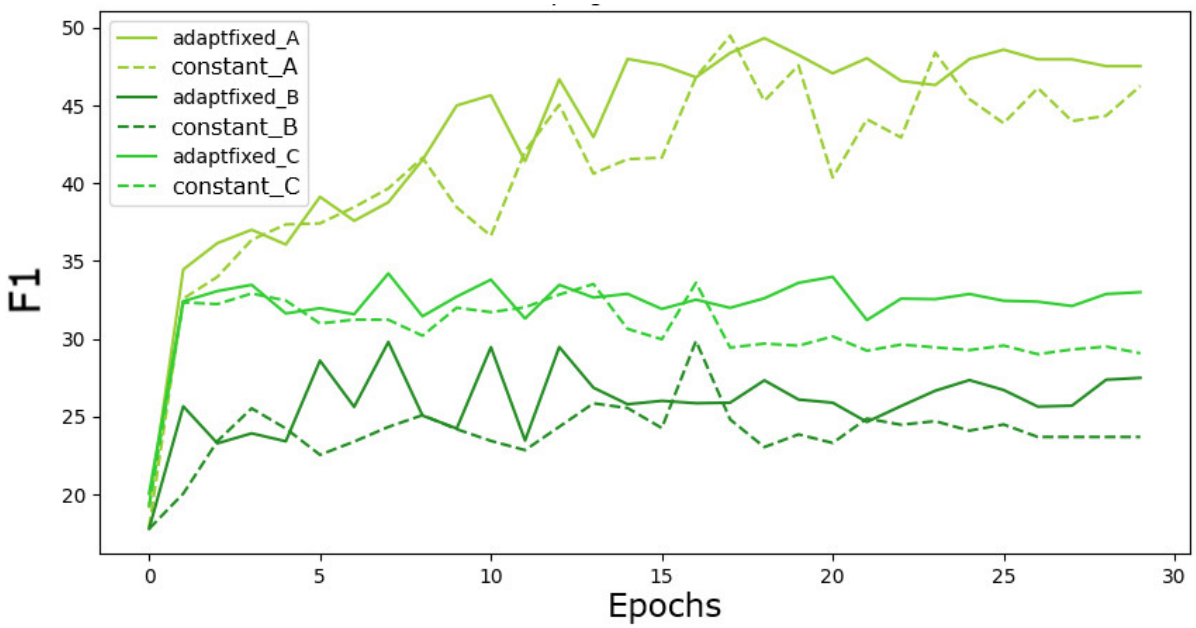

Figure 3: F1 score during training for 30 epochs for each task in CLPsych 19 for the Constant (dotted) and Adapt-Fixed (solid) models.

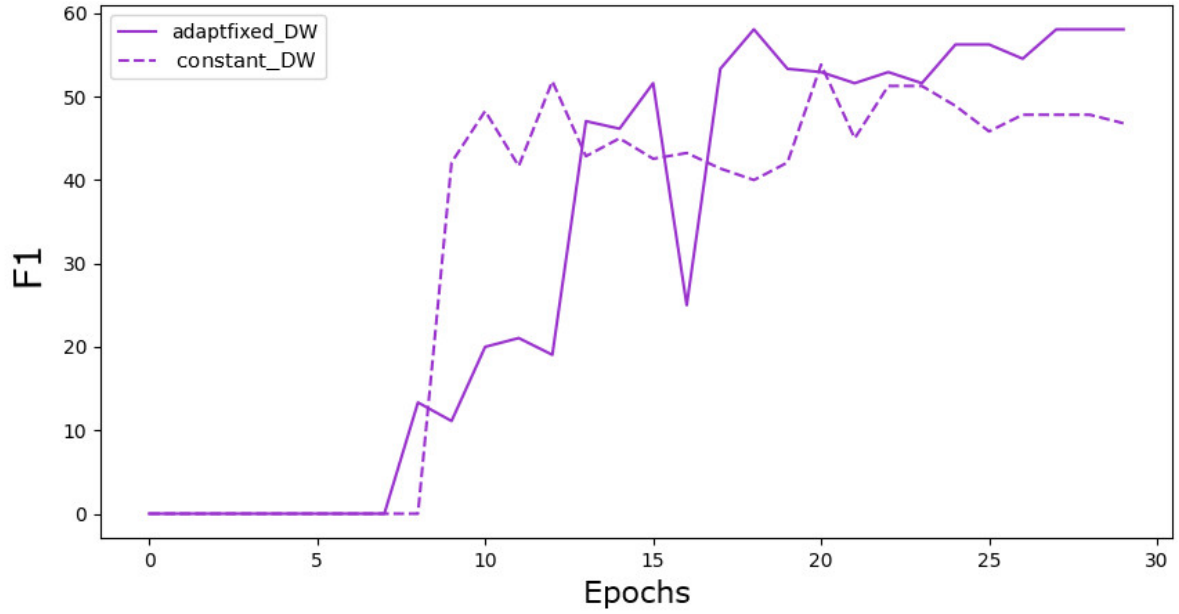

Figure 4: F1 score during training for 30 epochs for each task in DAIC-WOZ for the Constant (dotted) and Adapt-Fixed (solid) models.