Identifying Depressive Symptoms from Tweets: Figurative Language Enabled Multitask Learning Framework

WARNING: This paper contains examples which are depressive in nature.

Shweta Yadav[*], **Jainish Chauhan**[†], **Joy Prakash Sain**[‡], **Krishnaprasad Thirunarayan**[‡]
Amit Sheth[§], and **Jeremiah Schumm**[‡]
[*] LHNCBC, U.S. National Library of Medicine, MD, USA
[†]Indian Institute of Technology Gandhinagar, India
[‡] Wright State University, OH, USA
[§]University of South Carolina, SC, USA
[*]shweta.shweta@nih.gov, [†]chauhan.jainish@iitgn.ac.in, [§]AMIT@sc.edu
[‡]{sain.9,t.k.prasad,jeremiah.schumm}@wright.edu

Abstract

Existing studies on using social media for deriving mental health status of users focus on the depression detection task. However, for case management and referral to psychiatrists, healthcare workers require practical and scalable depressive disorder screening and triage system. This study aims to design and evaluate a decision support system (DSS) to *reliably* determine the depressive triage level by capturing *fine-grained depressive symptoms* expressed in user tweets through the emulation of Patient Health Questionnaire-9 (PHQ-9) that is routinely used in clinical practice. The reliable detection of depressive symptoms from tweets is challenging because the 280-character limit on tweets incentivizes the use of creative artifacts in the utterances and figurative usage contributes to effective expression. We propose a novel BERT based robust multi-task learning framework to accurately identify the depressive symptoms using the auxiliary task of figurative usage detection. Specifically, our proposed novel task sharing mechanism, *co-task aware attention*, enables automatic selection of optimal information across the BERT layers and tasks by soft-sharing of parameters. Our results show that modeling figurative usage can demonstrably improve the model's robustness and reliability for distinguishing the depression symptoms.

1 Introduction

The recent survey conducted by WHO shows that a total 322 million people in the world are living with depression. At its most severe, depression can lead to suicide and is responsible for 850, 000 deaths every year (WHO and others, 2017). Early detection and appropriate treatment can encourage remission and prevent relapse (Halfin, 2007). However, the stigma coupled with the depression makes patients reluctant to seek support or provide truthful answers to physicians (Haselton et al., 2015). Additionally, clinical diagnosis is dependent on the self-reports of the patient's behavior, which requires them to reflect and recall from the past, that may have obscured over time. In contrast, social media offers unique platform for people to share their experiences in the moment, express emotions and stress in their raw intensity, and seek social and emotional support for resilience. As such, the depression studies based on social media offer unique advantages over scheduled surveys or interviews (Coppersmith et al., 2014; De Choudhury and De, 2014; Manikonda and De Choudhury, 2017; De Choudhury et al., 2016). Social media self-narratives contain large amounts of implicit and reliable information expressed in real-time, that are essential for practitioners to glean and understand user's behavior outside of the controlled clinical environment. Majority of these existing studies have formulated the social media depression detection task as a binary classification problem (i.e., depressive/non-depressive) and therefore are limited to only identifying the depressive users.

Symptom	Sample Tweet
S1: Lack of Interest	*I don't think I care about anything at all lol it's f*** w my brain , boutta go nuts*
S2: Feeling Down	*im alone at home with no money and depressed as f***'*
S3: Sleeping Disorder	*This is not a good night at all . Rough .*
S4: Lack of Energy	*i have not moved all day . still in bed .*
S5: Eating Disorder	*its good not to eat..*
S6: Low Self-Esteem	*i am so ugly but will never stop posting pics 4 validation lol*
S7: Concentration Problem	*my mind is screaming so many things*
S8: Hyper/Lower Activity	*wish i didn't sit around every day wishing all my days away .why .*
S9: Self-Harm	*Cut all my elbow up but can't feel it*

Table 1: Sample tweets (rephrased) and their associated PHQ 9 symptoms.

To assist healthcare professionals (HPs) intervene in a timely manner such as with an automatic triaging, it is necessary to develop an intelligent decision support system that provides HPs fine-grained depression related symptoms. The triage process is a critical step in giving care to the patients because, by prioritizing patients at different triage levels based on the severity of their clinical condition, one can enhance the utilization of healthcare facilities and the efficacy of healthcare interventions. There have been a few efforts to create datasets for capturing depression severity, however they are limited to **(1)** only clinical interviews (Valstar et al., 2013; Ringeval et al., 2019; Gratch et al., 2014) and questionnaires (De Choudhury et al., 2013), and **(2)** individuals who voluntarily participate in the study (De Choudhury et al., 2014).

In this work, we exploit the Twitter data to identify the indications (specifically, `PHQ-9` guided symptoms) of depression. We developed a high quality dataset consisting of total $12,000$ tweets, with 3738 tweets posted by 205 self-reported depressed users over 2 weeks time, which were manually annotated using `PHQ-9` questionnaire (Kroenke and Spitzer, 2002) based symptoms categories. In Table-1, we provide sample tweets associated with the nine item `PHQ-9` depression symptoms. Our research hypothesis is that depressed individuals discuss their symptoms on Twitter that can be tracked reliably. Nonetheless, user social-media post offer unique challenges as discussed below:

- **Usage of the figurative language:** First, the depressive users often tend to use figurative language (*'FL'*) elements such as sarcasm and metaphor, to describe their symptoms. For example, one user wrote metaphorically, *"My skin is paper, razor is the pen"*, while another user wrote *"I want to cut myself"*. While both of these utterances refer to one specific medical concept "Self-Harm", the first sentence utilizes paper and pen metaphorically to convey self-harm. Furthermore, previous studies utilizing social media data reported prediction errors when drug or symptom names were utilized in a figurative sense (Iyer et al., 2019).
- **Usage of implicit sense:** The creative expressions used by depressive users also possess implicit sense not evident from a literal reading. For example, one user expresses their desperation as, *"What if life comes after death, grab my knife and find out myself."* implicitly referring to "Self-Harm", while another user gives a compliment with *"You have a killer look."*, or captures anger through *"If looks could kill, I would be dead by now."*. Other challenges include recognizing misspelled words, slangs, acronyms and unconventional contractions.
- **Usage of highly polysemous words:** The vocabulary of social media language offers polysemous words that require understanding of the context to determine the semantic labels. For example, "*woke up and nose started bleeding*" and "*I wish I had the nerve to press the blade deeper into my skin so I don't stop bleeding this time.*", use "bleeding" in different contexts and senses.

To account for this creative linguistic device widely observed in utterances of depressive users, we propose a **Fi**gurative **La**nguage enabled **M**ulti-**T**ask **L**earning framework (`FiLaMTL`) that works on the concept of task sharing mechanism (Ruder, 2017; Yadav et al., 2018b; Yadav et al., 2019). In this work, we improve the performance and robustness of the `FiLaMTL` for the primary task of *'symptom identification'* combined with the supervisory task *'figurative usage detection'* in a multi-task learning setting. We introduce a mechanism named *'co-task aware attention'* which enables the layer-specific soft sharing of the parameters for the tasks of interest. The proposed attention mechanism is parameterized with the

task-specific scaling factor for BERT (Devlin et al., 2019) layers. BERT enables even the low-resource tasks to benefit from deep bi-directional architectures and the unsupervised training framework to obtain the context-aware encoded representation. The virtue of this model is its ability to learn the task-specific representation of the input tweet by coordinating among the layers and between the tasks.
Contributions:

1. We propose a robust multi-task learning framework for identifying fine-grained `PHQ-9` defined symptoms from depressive tweets that takes into consideration the figurative language wired in the communication of depressive users. To the best of our knowledge, this is the first study in the depression domain that accounts for figurative language in the users social-media post.
2. We introduce an effective way to *fine-tune the BERT model* for multi-task learning using '*co-task aware attention*' to better encode the feature across the different BERT layers and tasks. This mechanism allows the model to learn the layer and task-specific parameters, to control the information flow from each BERT layer, in end-to-end model training.
3. To evaluate our study, we created a corpus of $12,155$ tweets, with 3738 tweets annotated with 9 `PHQ-9` symptom classes (validated by the collaborator psychiatrists). Additionally, these tweets were also labeled with the figurative classes: *metaphor* and *sarcasm*.

2 Related Works

Depending upon the data modalities and depressive markers, we categorize the existing literature as follows:

1. **Linguistic Marker:** Language often reflects how people think and is a well known tool used by psychiatrists to assess the mental health condition of the people (Fine, 2006). Numerous research (Coppersmith et al., 2014; De Choudhury et al., 2013; De Choudhury et al., 2014; Yadav et al., 2020b) has shown that modeling of word-use and social language combined with network analysis has been effective in recognizing depression. A widely adopted resource for understanding the linguistic patterns in mental health is the well-known Linguistic Inquiry Word Count (LIWC) (Pennebaker et al., 2007). Other researchers exploited sentiment analysis (Xue et al., 2014; Huang et al., 2014; Yadav et al., 2018a), topic modeling (Resnik et al., 2015) and emotion features (Chen et al., 2018; Aragón et al., 2019) to detect depression. Furthermore, substantial progress has been made with the introduction of a shared task (Coppersmith et al., 2015; Milne et al., 2016). Recently, most of the existing studies (Yates et al., 2017; Benton et al., 2017) have drifted from the traditional linguistic indicators to automated feature generation using the neural network based technique to predict or assess at-risk depressive users.
2. **Visual Marker:** Visual information such as head pose, body movement, facial expressions, gestures and eye blinks provide important cues in analyzing depression. Girard et al. (2014) examined if there exists a relationship between non-verbal cues and depression severity using Facial Action Coding System (Ekman and Friesen, 1978). In another prominent study utilizing FACS, Scherer et al. (2013) identified that a more downward gaze angle, dull smiles, shorter average lengths of smile, longer self-touches may predict depression. Several studies (de Melo et al., 2019; Cummins et al., 2013) have also investigated the Space-Time Interest Point (STIP) features that capture spatio-temporal changes such as facial motion and the movement in the hand, foot, and head.
3. **Speech Marker:** Recent studies have shown the potential for exploiting speech for depression detection and monitoring (Cummins et al., 2015a; Cummins et al., 2015b; Scherer et al., 2014). Numerous research (Mundt et al., 2012; Hönig et al., 2014) have revealed the strength of prosodic markers, specifically the *speech-rate* to analyze the level of depression. Moore II et al. (2007) proposed a depression classification system based on a wide range of acoustic feature like prosodic, spectral, voice quality, and glottal feature. Other prominent studies (Mundt et al., 2012; Cummins et al., 2011) have explored spectral features like prosodic timing measures, mel-frequency cepstral coefficients (MFCC) and glottal features to accurately classify depressed and control groups.
4. **Multi-modal Marker:** In recent times, there is visible surge in investigating multi-modal indicators to diagnose depression, particularly due to publicly available datasets made available through

research challenges like Audio/Visual Emotion Recognition (AVEC) Workshop Depression Sub-challenge (2013-2019) (Schuller et al., 2011; Valstar et al., 2013; Ringeval et al., 2019) and popular Distress Analysis Interview Corpus (DAIC) (Gratch et al., 2014). Several computational models (Tzirakis et al., 2017; Ringeval et al., 2017; Ringeval et al., 2018) based on machine learning and sophisticated deep learning techniques have been proposed to address the challenges posed by AVEC each year. The best system at AVEC 2019 (Ray et al., 2019) proposed an attention based fusion technique to judiciously select the feature representation obtained from multimodal source.

3 Corpus Creation and Analysis: D2S

In this section, we describe how we crawled our dataset **D**epression to (**2**) **S**ymptoms (named as D2S) using the Twitter streaming application programming interface, filtered out irrelevant profiles, annotated the tweets of depressive users and verified the annotations by a psychiatrist to prepare the gold standard dataset.

1. **Dataset Crawling:** We utilized the lexicon developed by (Yazdavar et al., 2017) in collaboration with a psychologist. The lexicon contains around 1000 depression-related terms categorized into nine categories of symptoms from `PHQ-9`. A subset of highly informative depression indicative terms from the lexicon, that are likely to be used by depressive individuals, was used as seed terms to crawl the public profiles of twitter users with at least one of those filtered terms in their profile description. Through this process, we collected $5,000$ users and their tweets.

2. **Filtering and Identifying Depressed Users:** As users on social media often use sarcasm and metaphor to implicitly express their feelings, contemporary approaches that do not capture context well, miss sub-population of depressive users. To improve upon these approaches, we proceed as follows. After filtering out the retweets, we removed the profiles with less than 100 tweets and obtained 1567 users. To emulate the `PHQ-9` using social media, we chunked the tweets of each profile into two week buckets. To ensure the high quality data and identify potential depressive profiles with severity level mild to severe based on `PHQ-9` scoring, we filtered the profiles based on their frequency of post. After filtering out the profiles who had not tweeted at least 5 days in the most recent bucket, we obtained 575 profiles. Although these profiles had depression-related terms in the description, due to the lack of context-sensitivity in the profile identification process, a subset of those were false positives, i.e., non-depressive. A few of these non-depressive profiles were meant to share motivational quotes for depressive users. We strictly examined the visual (i.e., profile image and shared images) and linguistic markers (i.e., profile name, description and tweets) of each of those profiles and removed the users having no depressive tweets. Finally, we obtained 205 depressive users and selected the bucket of the most recent tweets over two weeks for annotation, which sums to 3738 tweets.

3. **Anonymization, Annotation and Verification:** Prior to annotation, we anonymized the user profiles with random numbers and replaced the mentions and URLs in tweets with strings '@USER' and 'URL' respectively. Four native English speakers from multiple disciplines were assigned to independently annotate the tweets into 9 categories of `PHQ-9`. The annotators were also asked to identify the tweets having usage of *FL* such as sarcasm and metaphor. The annotators were provided with the definitions and samples of annotated tweets from each of those 9 categories of `PHQ-9` and as well as *FL*. The average inter-annotator reliability scores for the symptoms, depressive vs. non-depressive, and figurative classes were K=0.83, K=0.87, and K=0.79, respectively, based on Cohen's Kappa statistics. We resolved the conflicting annotations with the majority voting strategy and resolved the ones voted evenly by a psychiatrist. After preparing the final gold standard data, we randomly selected 100 annotated tweets from each of the symptom categories, including the non-depressive ones and verified by a psychiatrist.

4. **Data Analysis:** The final data[1] comprises 3738 depressive tweets, each tagged with the number of

[1] Limitation:(i) Only a sub-population (i.e., those with self-reported diagnosis) is identified by this method , (ii) Twitter users

Depressive														Non-depressive
	Symptoms										Figurative Language			
	S1	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**Total**	**Sarc**	**Meta**	**Total**	**Total**
#Tweets	494	657	261	301	473	1054	136	122	970	3738	668	1106	1485	8417
Avg.Len	13.02	14.13	13.03	13.08	16.08	12.83	13.31	15.63	12.33	12.98	13.17	13.41	13.24	10.96

Table 2: The statistics of depressive and non-depressive tweets. S[1-9]: PHQ-9 depressive symptoms, Sarc: Sarcasm, Meta: Metaphor, #Tweets: No. of tweets per class, Avg.Len: Average length of tweets.

Symptom	Topics of Interest
S1	*bored, disgusting, sick of, tired of it, dont want to, so f** miserable, tired of being*
S2	*depressed, alone, isolate, given up, no friend, cant deal, want to talk, in my room*
S3	*awake, sleepless, nightmares, insomnia, cant sleep, wish sleep, up all night, body is begging*
S4	*exhausted, tired, weak, my energy, dont have energy, tired to look, feel myself falling*
S5	*binge, fasting, eating disorder, eat again, always eating, forced to eat, am eating ?*
S6	*failure, ugly, worthless, hate myself, fat piece, self hatred, piece of sh**, feel like trash*
S7	*thoughts, confused, overthinking, brain, my head, am losing, losing mind, my mind off*
S8	*quiet, slowly, attention, nervous, social anxiety, dead quiet, dont wanna move*
S9	*cut, hang, blade, die, suicidal, rip skin, suicide attempt, car hit, kill myself, of the road*

Table 3: Sample of Topics identified from depressive tweets. S[1-9]: PHQ-9 symptoms

symptoms (single or multiple), and the presence of figurative expressions (either or both) it exhibits, and 8417 non-depressive tweets. Out of these depressive tweets, 1485 ($\sim 40\%$) tweets use *FL.* = We performed topic analysis to examine the usage of utterances associated with each symptom. Table 3 illustrates the topic distribution of each symptom. We observe from the table, to express their feelings, the depressive individuals use metaphoric phrases such as *'body is begging'*, and *'feel like trash'*; sarcastic expressions such as *'am eating ?'*; implicit utterances such as *'up all night'*, and *'feel myself falling'*.

5. **Ethical Concerns:** Psychiatric research using social media data poses several ethical concerns regarding user privacy, which should necessarily be taken into consideration (Hovy and Spruit, 2016; Valdez and Keim-Malpass, 2019). Following the ethical practices, as adopted by the previous research on Twitter data (Coppersmith et al., 2015), we constructed our dataset using only public twitter profiles. We anonymized the profiles before presenting it to the annotators who pledged not to make attempts to contact or deanonymize any of the users or share the data with others. The dataset will be shared with researchers who agree to follow the similar ethical guidelines.

4 Methods

Our proposed approach to identify the depressive symptoms, is assisted by the Bidirectional Representation from Transformers (BERT) and multi-task learning (Yadav et al., 2020a) with soft-parameter sharing. This section describes the proposed methodology for identifying the depression symptoms from user tweets.

4.1 Problem Definition

Given an input tweet sequence T consisting of n words, i.e., $T = \{w_1, w_2 \dots w_n\}$, our multi-label classification task is to learn the function $fun(.)$ that predicts the set of probable classes $\bar{y_1}, \bar{y_2}, \dots, \bar{y_k}$ from the set of class labels, Y. Mathematically,

$$\bar{y_1}, \bar{y_2}, \dots, \bar{y_k} = fun(T, \theta_1, \theta_2, \dots, \theta_P) \tag{1}$$

where, $\theta_i, (i = 1, \dots, P)$ is the model parameter. The function $fun(.)$ returns the probability of each symptom class assigned to the tweet. We choose the set of best probable class based on the particular threshold value, a hyper-parameter. In our proposed multi-task learning framework, the primary task is symptom identification with nine labels from PHQ-9. We consider the figurative usage detection as the auxiliary task having three class labels: '*metaphor*','*sarcasm*', and '*others*'.

cannot be reflective of the entire population, and (iii) it cannot be verified if self-reported depressed users are being truthful

4.2 Background

BERT is one of the powerful language representation models that has the ability to make predictions that go beyond the natural sentence boundaries (Lin et al., 2019). Unlike CNN/LSTM model, language models benefit from the abundant knowledge from pre-training using self-supervision and have strong feature extraction capability. It uses word-piece tokenizer (Wu et al., 2016) to tokenize the input sentence. When the model uses word-piece token and randomly mask a portion of the word to predict in the masked language model (MLM) task then the model attempts to recover a piece of the word rather than the whole word. To mitigate this issue, recently, Devlin et al. (2019) released an updated version of BERT, which is called Whole Word Masking (`wwm`). We use the pre-trained `wwm` BERT model[2] having 24 Transformer layers (L), each having 16 heads for self-attention and hidden dimension of 1024. The input to the BERT model is the tweet $T = \{w_1, w_2, \ldots, w_n\}$. It returns the hidden state representation of each input word from each Transformer layer. Formally,

$$\begin{gathered} T_1, T_2, \ldots, T_L = BERT([w_1, w_2, \ldots, w_n]) \\ \text{where} \quad s_1^i, s_2^i, \ldots, s_n^i = T_i \quad \text{and} \quad h_1, h_2, \ldots, h_L = s_1^i, s_1^2, \ldots, s_1^L \end{gathered} \tag{2}$$

where, s_i^j is the i^{th} token representation obtained from j^{th} transformer encoder, and $h_i \in \mathbb{R}^d$ and d is the dimension of the `[CLS]` token hidden state representation obtained from BERT.

4.3 Figurative Language enabled Multi-task Learning (`FiLaMTL`) Framework

We explore the utility of learning two tasks together in a `FiLaMTL` framework. For depression symptom identification (SI) task, `FiLaMTL` helps achieve inductive transfer from figurative usage detection (FUD) task by leveraging additional sources of information to improve performance on the primary task. We focus on designing the soft-parameter sharing rather than hard-parameter as it offers a way to effectively share the required parameters between the tasks (Misra et al., 2016). We achieve this with *co-task aware attention* module that finds the best shared representation for multi-task learning. Specifically, the proposed network models shared representations using linear combinations, and learns the optimal combinations for the primary and the auxiliary tasks. Let us denote the hidden states (from eq. 2) for primary task (SI), , as $H_s \in \mathbb{R}^{L \times d}$ and the auxiliary task (FUD), as $H_f \in \mathbb{R}^{L \times d}$. For a given layer $l \in L$, we compute the effective shared representation as follows:

$$\begin{aligned} r_s^l &= \alpha_{(s,s)} \times \beta_{(l,s,s)} \times h_s^l + \alpha_{(s,f)} \times \beta_{(l,s,f)} \times h_f^l \\ r_f^l &= \alpha_{(f,f)} \times \beta_{(l,f,f)} \times h_s^l + \alpha_{(f,s)} \times \beta_{(l,f,s)} \times h_f^l \end{aligned} \tag{3}$$

where $h_s^l \in \mathbb{R}^d$ and $h_f^l \in \mathbb{R}^d$ are the hidden state representations obtained from l^{th} BERT layer for SI and FUD respectively. $r_s^l \in \mathbb{R}^d$ and $r_f^l \in \mathbb{R}^d$ are modified hidden state representation of l^{th} BERT layer, after applying the effective soft-sharing of features across the two tasks. We will discuss the scaling factors α and β shortly.

4.3.1 Soft-parameter Sharing between Tasks

In multi-task learning, inductive bias of auxiliary task helps to improve the performance of primary task. However, the parameter sharing between the tasks is non-trivial, as we need an optimal strategy to improve the performance of primary task. Towards this end, we devise a strategy to automatically learn the factor by which a feature from a particular task need to be accommodated for learning the optimal set of shared features for a given task. This co-task aware sharing of the features leads to the optimal linear combination of feature spaces across the task. Given the two tasks: "*symptom identification*" and "*figurative usage detection*", we learn how much of each task's features contribute to form the shared feature space, which leads to the overall improvement of the tasks. We achieve this by introducing a "*co-task factor matrix*" $\alpha \in \mathbb{R}^{T \times T}$, where T is the number of tasks at hand. In our case $T = 2$, as we are dealing with two tasks here. An element $\alpha_{(x,y)}$ of the matrix α denotes "*the factor of which y^{th} task feature obtained from a particular layer of BERT should contribute to the shared-feature representation*

[2] `https://bit.ly/3eQAZSY`

for x^{th} task". Moreover, this matrix is learned by end-to-end training of the proposed multi-task learning framework.

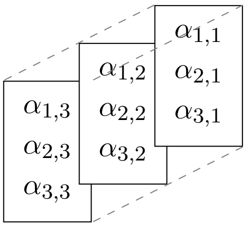

Figure 1: Representation of co-task factor matrix for the three tasks.

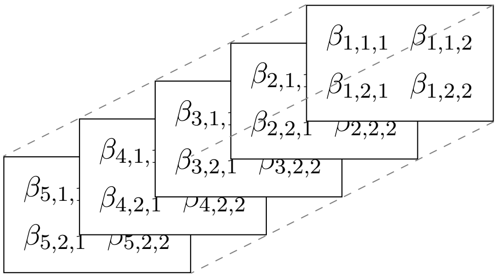

Figure 2: Representation of layer factor matrix for five layers and two tasks.

4.3.2 Soft-parameter Sharing between Layers

Given input tokens and task t, BERT produces the set of layer activation $h_t^1, h_t^2, \ldots, h_t^L$ in the form of hidden state representation. Each layer of the BERT attempts to address certain problem (Tenney et al., 2019) and feed the information to the upper layer. Tenney et al. (2019) discovered that information learned at a few layers of BERT is sufficient to reliably model and address the lower-level tasks of an NLP pipeline such as parts-of-speech tagging, but, to model the higher level tasks such as relation extraction or co-reference resolution, we need many layers. The hidden state representation $h_{t_1}^l$, obtained from the l^{th} layer of BERT for a given task t_1, may not be as useful for another task t_2. It always depends on task complexity. Inspired by this, we introduce the soft-parameter sharing among the BERT layers. Similar to the soft-parameter sharing between tasks, we propose a mechanism to automatically learn the factors by which a feature from a particular BERT layer needs to be accommodated for learning the optimal set of shared features for a given task. We achieve this by introducing a "*layer factor matrix*" $\beta \in \mathbb{R}^{L\times T\times T}$, where L and T denote the number of BERT layers and the number of tasks respectively. An element $\beta_{(x,y,z)}$ of the matrix β denotes "*the factor of which z^{th} task feature obtained from x^{th} layer of BERT should contribute to the shared-feature representation for y^{th} task*". Similar to the co-task factor matrix α, the layer factor matrix β is also a network parameter and can be learned by end-to-end training of proposed multi-task learning framework. We shown the hypothetical matrices for α and β in Fig 1 and 2 respectively.

4.4 Symptom Identification and Figurative Usage Detection

For each task, we obtained the effective shared-feature representation from each BERT layer. The final feature is obtained by the average pooling of each individual feature as follows:

$$z_s = \frac{1}{L}\sum_{i=1}^{i=L} f(\mathbf{W_s}.r_s^i + \mathbf{b_s}) \quad \text{and} \quad z_f = \frac{1}{L}\sum_{i=1}^{i=L} f(\mathbf{W_f}.r_f^i + \mathbf{b_f}) \tag{4}$$

where $z_s \in \mathbb{R}^d$ and $z_f \in \mathbb{R}^d$ correspond to the final pooled features for the task symptom identification and figurative usage detection respectively. $\mathbf{W_s}$, $\mathbf{W_f}$, $\mathbf{b_s}$ and $\mathbf{b_f}$ are the weight and bias matrices and f is a non-linear activation. For symptom identification, we employ a feed-forward network with sigmoid activation function to find the probability of a class label belonging to a given tweet,

$$l_s = sigmoid(\mathbf{W_{sl}}.z_s + \mathbf{b_{sl}}) \quad \text{and} \quad l_f = sigmoid(\mathbf{W_{fl}}.z_f + \mathbf{b_{fl}}) \tag{5}$$

where l_s and l_f are logits for the symptom identification and figurative usage detection tasks respectively. $\mathbf{W_{sl}}$, $\mathbf{W_{fl}}$, $\mathbf{b_{sl}}$ and $\mathbf{b_{fl}}$ are the weight and bias metrices.

5 Experiments

We will first provide detail about baseline models and then present our results on SI and FUD task. Later, we will assess the performance of our approach on depression detection task.

5.1 Implementation Details

We shuffle the D2S dataset and split it into 70% training (TRAIN), 10% development (DEV), and 20% test (TEST). For both symptom identification (SI) and figurative usage detection (FUD) models, we have chosen the hyper-parameters using the development set. In all of our experiments, we have fine-tuned the BERT-wwm model for 10 epochs with the batch size of 32. We fine-tune and extracted the features from top three layers of the BERT model. In the proposed FiLaMTL framework, the overall loss of the network is the weighted factor of the loss computed for both the tasks. The network is trained with the binary-cross entropy loss function for both tasks. We set the weight 0.7 for symptom identification and 0.3 for figurative usage detection task. We use sigmoid activation function as the non-linear activation to project the BERT hidden state representation to another representation of dimension 256. We used Adam optimizer (Kingma and Ba, 2014) with a fixed learning rate of 0.0001. For regularization, we used dropout (Srivastava et al., 2014) with a value of 0.5 on each of the hidden layers. We then ran each best model on TEST, and report recall, precision, and F1-Score.

5.2 Baseline Models and Results

We compare the FiLaMTL with the highly competitive baseline models and evaluated the model on Precision, Recall, and weighted F1-Score. Since BERT has already demonstrated remarkable performance on multiple NLP tasks over SOTA deep learning (DL) methods, we restricted ourselves to using BERT over DL techniques as our baseline model discussed below:
(1) *STL-BERT*: This is a domain-adapted BERT-based model proposed for the SI and FUD tasks, which fine tuned the BERT model on corresponding dataset.
(2) *MTL-H-BERT*, Dense: A variation of the multi-task BERT model where a single BERT model generate the features for both the tasks. The features are transformed to another representation by the task-specific dense layer.
(3) *MTL-S-BERT, Cross-Stitch*: This model is the re-implementation of Misra et al. (2016), where the model learns an optimal combination of shared and task-specific representations using soft-parameter sharing via cross-stitch units.
(4) *MTL-S-BERT, Co-Attention*: This model was inspired by the framework of Lu et al. (2016). Firstly, we compute the word-level attention weight as discussed in Lu et al. (2016) for the hidden state representation of both the tasks. These weights were multiplied with the corresponding hidden state representation to compute the attentive features. Similar to ***MTL-S-BERT, Cross-Stitch***, we employed the cross-stitch units to obtained the final hidden state representation for both the tasks.

Table-4 provides a comparative summary of the results of our proposed approach over the baseline models demonstrating that our '*co-task-aware*' multitask FiLaMTL model outperforms the SOTA single task learning model and the variations of BERT inspired multi-task learning models. Basically, we train MTL model in two different ways: *hard-parameter sharing* and *soft-parameter sharing*. We can visualize from Table-4, the multitask learning framework based on the soft-parameter sharing (MTL-S-BERT, Cross-Stitch) can assist the performance of the main task as well as in the FUD task over the single task model. However, multi-task hard parameter sharing model (MTL-H-BERT, Dense) was found to be useful only in the FUD over SI task. This may be due to the noise in the dataset over the tasks, which prevents to learn task-specific efficient representation required to correctly identify the symptom from the input tweet. We also observe that soft-parameter sharing based baseline model (MTL-S-BERT, Co-Attention) could not produce the desired results, because of the additional attention mechanism over the strong self-attention mechanism, overfits the model which leads to the degradation in the performance. The existing strategy of the parameter-sharing shows the inconsistency in the performance of the multi-task learning framework. We exploited variation of soft-parameter sharing to further understand the relevance of *co-task aware* attention in the multi-task learning setting. Our FiLaMTL outperforms both the hard-parameters and soft-parameters sharing based baseline models on both the tasks (Table 4). This also demonstrates that providing information about *FL* to the BERT model significantly improves the performance of the model and thus enabling generalization to other tasks related to text classification where there is extensive usage of *FL*.

Models	Symptom Identification (SI)			Figurative Usage Detection (FUD)		
	P	R	F1	P	R	F1
STL-BERT	74.64	71.46	73.02	63.32	71.35	67.09
MTL-H-BERT, Dense	73.76	72.15	72.95	62.97	75.17	68.53
MTL-S-BERT, Cross-Stitch (Misra et al., 2016)	76.17	71.00	73.50	68.48	77.35	72.65
MTL-S-BERT, Co-Attention (Lu et al., 2016)	74.09	70.43	72.21	69.28	78.44	73.58
FiLaMTL	**76.46**	**73.65**	**75.03**	**75.67**	**75.44**	**75.55**

Table 4: Performance comparisons of our proposed approach with baselines for identifying depressive symptoms and figurative usage. The results are reported for only positive classes.

Dataset	Model	Accuracy
D2S	BERT	96.91
	FiLaMTL-*fine-tuned*	97.44
CLPsych	BERT	70.01
	FiLaMTL-*fine-tuned*	70.79

Table 5: Evaluation of FiLaMTL on our D2S dataset and CLPsych 2015 dataset for depression detection task.

No.	Control	Depressed	Total
# Users in train	151	69	220
# Users in test	300	150	450
# Tweets in train	254786	110585	365371
# Tweets in test	541282	284777	826059

Table 6: Dataset Statistics on the CLPsych 2015 shared task corpus used in our experiments for depression detection.

Evaluating the FiLaMTL on Depression Detection Task: To further verify the effectiveness of our proposed FiLaMTL model, we utilized a transfer learning procedure, where an intermediate shared model obtained on the *SI* and *FUD* task is fine-tuned on the depression detection (DD) task. Towards that, we experimented with **(a)** STL-BERT and **(b)** FiLaMTL- fine-tuned for DD task, on D2S corpus and the bechmark CLPsych dataset (Coppersmith et al., 2015). The data statistics can be viewed in Table-6. In FiLaMTL (fine-tuned), we initialize the parameters of the BERT model with the obtained weighted from the FiLaMTL model (BERT model of FUD task) reported in Table 4. Our motivation is that first fine-tuning on the *FUD* and *SI* task can assist the LMs to adapt to the depression domain with some understanding of figurative usage detection, thus making the training on DD more stable. Table-5, summarizes the results on DD task, after the transfer learning procedure. It can be noticed that fine-tuning FiLaMTL model improves the performance over vanilla BERT model on both the datasets [3]. This proves that FiLaMTL can be generalized for other tasks related to biomedical NLP task where there is extensive usage of *FL*.

Analysis: To get a deeper insight into how FiLaMTL performed over the baseline models, we examined the classification of tweets on SI task and came up with the following observations:

1. **Understanding figurative sense:** Table-7 shows the capability of our model to capture sarcastic and metaphoric senses in the utterances of depressive users. Our model performs better than STL-BERT in handling figurative tweets. The main reason why STL-BERT model misclassifies sarcastic or metaphoric tweets is because BERT has been trained on BookCorpus and Wikipedia corpus which has fewer *FL* fragments compared to that in social media.

2. **Understanding implication:** For the tweets where depression was implicit (cf., Table-7), most of the baseline models including MTL-series were prone to misclassification. As BERT based models tend to capture only local information available in a tweet, it fails to understand the implicit context of the subject. For example, if a tweet contains a keyword '*sleep*', the model will likely classify it as belonging to the PHQ class 3 related to sleeping disorder, without necessarily determining a different contextual use (such as "*permanent land of nod*" and "*going to sleep early tonight*"). However, our model with the *co-task aware attention* information sharing unit tends to have better coverage for identifying the depressive symptoms.

[3] We were not able to compare FiLaMTL over existing model developed on CLPsych dataset, due to: (1) different dataset statistics of what we obtained from organizers and the dataset used by the participants and (2) multiple ways of evaluations.

Case	Tweets	Actual Labels	Predicted Label				
			STL-BERT	MTL-H-BERT	MTL-S-BERT, Cross-Stitch	MTL-S-BERT, Co-attention	FiLaMTL
Understanding figurative sense	T1: holy sh**. i look like death	Low Self-Esteem	Low Self-Esteem, Self-Harm	Low Self-Esteem, Self-Harm	Low Self-Esteem	Self-Harm	Low Self-Esteem
	T2: hang on a rope or bated breath, whichever you prefer	Self Harm	Feeling Down	Feeling Down, Lack of Interest	Feeling Down	Self-Harm, Feeling Down	Self-Harm
Understanding implication	T3: i want a zombie to read my nutrition label, be happy to see it say low fat , and then eat me	Low Self-Esteem	Eating Disorder	Eating Disorder, Self-Harm	Eating Disorder, Low Self-Esteem	Eating Disorder	Low Self-Esteem
	T4: insomia can i pls sleep forever	Sleeping Disorder, Self-Harm	Sleeping Disorder	Sleeping Disorder	Sleeping Disorder	Sleeping Disorder, Self-Harm	Sleeping Disorder, Self-Harm

Table 7: Qualitative analysis of our proposed model, FiLaMTL, with the baseline models

Error Types	Tweets	Actual Labels	Predicted Label
Ambiguity	*T1: sorry i was such a failure*	Low Self-Esteem	Low Self-Esteem, *Feeling Down*
	T2: its worth noting that im not worth noting	Low Self-Esteem	Low Self-Esteem, *Feeling Down*
Cryptic tweets	*T3: shut up stomach*	Eating Disorder	*None*
	T4: im done	Self-Harm	*Feeling Down*
Multiple Symptoms	*T5: it really sucks being strong all the time. its so draining and when youre all depleted it feels like youre underwater.*	Lack of Interest, Lack of Energy, Concentration Problem, Hyper/Lower Activity	*Lack of Energy*, Hyper/Lower Activity
	T6: im 24/7. online. bored. hungry . sleepy	Lack of Interest, Sleeping Disorder, Eating Disorder	*Feeling Down*, *Lack of Energy*, Sleeping Disorder

Table 8: Exemplar description showing prominent errors made by our proposed approach.

Tweets	Actual Label	Predicted Label
T1: hang on a rope or bated breath, whichever you prefer.	Sarcasm	Sarcasm
T2: If looks could kill, I would be dead by now.	Sarcasm	Sarcasm
T3: people treat me like a god. they ignore my existence until they need something from me .	Metaphor	*Sarcasm*, Metaphor
T4: i dont see a future	Sarcasm	*Others*

Table 9: Qualitative analysis of FiLaMTL in identifying figurative language.

Error Analysis: Following are the major errors made by our approach on SI task:

1. **Ambiguity:** PHQ-9 classes related to sleeping disorder, eating disorder, and self-harm are easy to distinguish. However classes such as PHQ-1 (feeling down) and PHQ-6 (low self-esteem) are difficult to separate because of overlapping expressions, often leading to misclassification. As observed in Table-8, both these classes are semantically similar, challenging manual labelling.
2. **Cryptic tweets:** Our model is unable to handle tweets that are only a few words long. The lack of context required for robust identification of symptoms can only be remedied by consulting past user interactions and communications.
3. **Multiple Symptoms:** The FiLaMTL is unable to predict all the PHQ-9 classes indicated by a tweet leading to incompleteness as shown in Table-8, tweet T5 and T6.

6 Conclusion

In this research, we explored a new dimension of social media in Twitter to identify depressive symptoms. Towards this end, we created a new benchmark dataset (D2S) for identifying *fine-grained PHQ-9 emulated depressive symptoms* that contains figurative language. We also introduce a robust BERT based MTL framework that jointly learns to automatically discover complementary features required to identify the symptoms with the help of the auxiliary task of *figurative usage detection*. Our experimental results convincingly show the effectiveness of introducing figurative usage detection for depressive symptoms identification. In future, we aim to enhance the dataset with the other modalities like image and memes to assist the model in better understanding of figurative sense in symptom identification.

References

Mario Ezra Aragón, Adrian Pastor López-Monroy, Luis Carlos González-Gurrola, and Manuel Montes. 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.

Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017. Multi-task learning for mental health using social media text. *arXiv preprint arXiv:1712.03538*.

Xuetong Chen, Martin D Sykora, Thomas W Jackson, and Suzanne Elayan. 2018. What about mood swings: identifying depression on twitter with temporal measures of emotions. In *Companion Proceedings of the The Web Conference 2018*, pages 1653–1660.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60.

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.

Nicholas Cummins, Julien Epps, Michael Breakspear, and Roland Goecke. 2011. An investigation of depressed speech detection: Features and normalization. In *Twelfth Annual Conference of the International Speech Communication Association*.

Nicholas Cummins, Jyoti Joshi, Abhinav Dhall, Vidhyasaharan Sethu, Roland Goecke, and Julien Epps. 2013. Diagnosis of depression by behavioural signals: a multimodal approach. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 11–20.

Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F Quatieri. 2015a. A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71:10–49.

Nicholas Cummins, Vidhyasaharan Sethu, Julien Epps, Sebastian Schnieder, and Jarek Krajewski. 2015b. Analysis of acoustic space variability in speech affected by depression. *Speech Communication*, 75:27–49.

Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *ICWSM*.

Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3267–3276. ACM.

Munmun De Choudhury, Scott Counts, Eric J Horvitz, and Aaron Hoff. 2014. Characterizing and predicting postpartum depression from shared facebook data. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 626–638.

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 2098–2110. ACM.

Wheidima Carneiro de Melo, Eric Granger, and Abdenour Hadid. 2019. Combining global and local convolutional 3d networks for detecting depression from facial expressions. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–8. IEEE.

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, June. Association for Computational Linguistics.

P Ekman and W Friesen. 1978. Facial action coding system (facs): A technique for the measurement of facial action, palo alto, ca: Consulting.

Jonathan Fine. 2006. *Language in psychiatry: A handbook of clinical practice*. Equinox London.

Jeffrey M Girard, Jeffrey F Cohn, Mohammad H Mahoor, S Mohammad Mavadati, Zakia Hammal, and Dean P Rosenwald. 2014. Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses. *Image and vision computing*, 32(10):641–647.

Jonathan Gratch, Ron Artstein, Gale M Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, et al. 2014. The distress analysis interview corpus of human and computer interviews. In *LREC*, pages 3123–3128. Citeseer.

Aron Halfin. 2007. Depression: the benefits of early and appropriate treatment. *The American journal of managed care*, 13(4 Suppl):S92–7.

Martie G Haselton, Daniel Nettle, and Damian R Murray. 2015. The evolution of cognitive bias. *The handbook of evolutionary psychology*, pages 1–20.

Florian Hönig, Anton Batliner, Elmar Nöth, Sebastian Schnieder, and Jarek Krajewski. 2014. Automatic modelling of depressed speech: relevant features and relevance of gender. In *Fifteenth Annual Conference of the International Speech Communication Association*.

Dirk Hovy and Shannon L Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598.

Xiaolei Huang, Lei Zhang, David Chiu, Tianli Liu, Xin Li, and Tingshao Zhu. 2014. Detecting suicidal ideation in chinese microblogs with psychological lexicons. In *2014 IEEE 11th Intl Conf on Ubiquitous Intelligence and Computing and 2014 IEEE 11th Intl Conf on Autonomic and Trusted Computing and 2014 IEEE 14th Intl Conf on Scalable Computing and Communications and Its Associated Workshops*, pages 844–849. IEEE.

Adith Iyer, Aditya Joshi, Sarvnaz Karimi, Ross Sparks, and Cecile Paris. 2019. Figurative usage detection of symptom words to improve personal health mention detection. *arXiv preprint arXiv:1906.05466*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kurt Kroenke and Robert L Spitzer. 2002. The phq-9: a new depression diagnostic and severity measure. *Psychiatric annals*, 32(9):509–515.

Chen Lin, Timothy Miller, Dmitriy Dligach, Steven Bethard, and Guergana Savova. 2019. A bert-based universal model for both within-and cross-sentence clinical temporal relation extraction. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 65–71.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances in neural information processing systems*, pages 289–297.

Lydia Manikonda and Munmun De Choudhury. 2017. Modeling and understanding visual attributes of mental health disclosures in social media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 170–181. ACM.

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.

Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. 2016. Cross-stitch networks for multitask learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3994–4003.

Elliot Moore II, Mark A Clements, John W Peifer, and Lydia Weisser. 2007. Critical analysis of the impact of glottal features in the classification of clinical depression in speech. *IEEE transactions on biomedical engineering*, 55(1):96–107.

James C Mundt, Adam P Vogel, Douglas E Feltner, and William R Lenderking. 2012. Vocal acoustic biomarkers of depression severity and treatment response. *Biological psychiatry*, 72(7):580–587.

James W Pennebaker, Roger J Booth, and Martha E Francis. 2007. Linguistic inquiry and word count: Liwc [computer software]. *Austin, TX: liwc. net*, 135.

Anupama Ray, Siddharth Kumar, Rutvik Reddy, Prerana Mukherjee, and Ritu Garg. 2019. Multi-level attention network using text, audio and video for depression prediction. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, pages 81–88.

Philip Resnik, William Armstrong, Leonardo Claudino, Thang Nguyen, Viet-An Nguyen, and Jordan Boyd-Graber. 2015. Beyond lda: exploring supervised topic modeling for depression-related language in twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 99–107.

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*, pages 3–9.

Fabien Ringeval, Björn Schuller, Michel Valstar, Roddy Cowie, Heysem Kaya, Maximilian Schmitt, Shahin Amiriparian, Nicholas Cummins, Denis Lalanne, Adrien Michaud, et al. 2018. Avec 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition. In *Proceedings of the 2018 on audio/visual emotion challenge and workshop*, pages 3–13.

Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, et al. 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*, pages 3–12.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Stefan Scherer, Giota Stratou, Marwa Mahmoud, Jill Boberg, Jonathan Gratch, Albert Rizzo, and Louis-Philippe Morency. 2013. Automatic behavior descriptors for psychological disorder analysis. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8. IEEE.

Stefan Scherer, Giota Stratou, Gale Lucas, Marwa Mahmoud, Jill Boberg, Jonathan Gratch, Louis-Philippe Morency, et al. 2014. Automatic audiovisual behavior descriptors for psychological disorder analysis. *Image and Vision Computing*, 32(10):648–658.

Björn Schuller, Michel Valstar, Florian Eyben, Gary McKeown, Roddy Cowie, and Maja Pantic. 2011. Avec 2011–the first international audio/visual emotion challenge. In *International Conference on Affective Computing and Intelligent Interaction*, pages 415–424. Springer.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy, July. Association for Computational Linguistics.

Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou. 2017. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1301–1309.

Rupa Valdez and Jessica Keim-Malpass. 2019. Ethics in health research using social media. In *Social Web and Health Research*, pages 259–269. Springer.

Michel Valstar, Björn Schuller, Kirsty Smith, Florian Eyben, Bihan Jiang, Sanjay Bilakhia, Sebastian Schnieder, Roddy Cowie, and Maja Pantic. 2013. Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 3–10.

WHO et al. 2017. Depression and other common mental disorders: global health estimates.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Yuanyuan Xue, Qi Li, Li Jin, Ling Feng, David A Clifton, and Gari D Clifford. 2014. Detecting adolescent psychological pressures from micro-blog. In *International Conference on Health Information Science*, pages 83–94. Springer.

Shweta Yadav, Asif Ekbal, Sriparna Saha, and Pushpak Bhattacharyya. 2018a. Medical sentiment analysis using social media: towards building a patient assisted system. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Shweta Yadav, Asif Ekbal, Sriparna Saha, Pushpak Bhattacharyya, and Amit Sheth. 2018b. Multi-task learning framework for mining crowd intelligence towards clinical treatment.

Shweta Yadav, Asif Ekbal, Sriparna Saha, and Pushpak Bhattacharyya. 2019. A unified multi-task adversarial learning framework for pharmacovigilance mining. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5234–5245.

Shweta Yadav, Srivastsa Ramesh, Sriparna Saha, and Asif Ekbal. 2020a. Relation extraction from biomedical and clinical text: Unified multitask learning framework. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

Shweta Yadav, Joy Prakash Sain, Amit Sheth, Asif Ekbal, Sriparna Saha, and Pushpak Bhattacharyya. 2020b. Assessing the severity of health states based on social media posts. *arXiv preprint arXiv:2009.09600*.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. *arXiv preprint arXiv:1709.01848*.

Amir Hossein Yazdavar, Hussein S Al-Olimat, Monireh Ebrahimi, Goonmeet Bajaj, Tanvi Banerjee, Krishnaprasad Thirunarayan, Jyotishman Pathak, and Amit Sheth. 2017. Semi-supervised approach to monitoring clinical depressive symptoms in social media. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 1191–1198. ACM.