More than Just a Diagnosis: A Multi-Task Approach to Analyzing Bipolar Disorder on Reddit via DeMHeM

Rocco Zhang[†*], Shailik Sarkar[‡*], Abdulaziz Alhamadani[‡], Chang-Tien Lu[‡]
[*] Authors with equal contribution
[†] Thomas Jefferson High School for Science and Technology, Alexandria, VA 22312 USA roccozxyt@gmail.com
[‡] Department of Computer Science, Virginia Tech, Falls Church, VA 22043 USA {shailik, hamdani, ctlu}@vt.edu

2023 IEEE International Conference on Big Data (BigData) | 979-8-3503-2445-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/BigData59044.2023.10386934

***Abstract*—Mental health conditions affect millions of people today. While existing work on predicting mental health conditions from social media text focuses largely on depression and similar conditions, other less prominent disorders like bipolar tend not to receive in-depth analysis. Furthermore, these works tend not to analyze or model the correlated nature of these different disorders and conditions. To account for the coexistence and correlation of multiple mental health conditions, this paper introduces DeMHeM, a novel multitask framework designed for the descriptive classification of bipolar and related mental health topics on online platforms like Reddit. By treating each mental health category as a separate task, DeMHeM leverages both the shared latent and task-specific semantic feature space by integrating sentence-level and topic-level embeddings. It further incorporates Focal Loss for joint learning, inter-task parameter sharing, and regularization decay to optimize the prediction for the naturally skewed imbalanced dataset. Hence, the model distinguishes between different mental health categories and also models the correlation among them by categorizing each post into potentially multiple mental health categories. Next, we focus on a more insightful analysis by leveraging the predicted outcome of the model to study how the discussions differ based on the type and coexistence of different mental disorders. We analyze the entirety of the "r/bipolar" subreddit by applying our trained model to predict a category and then implementing keyword extraction techniques on each predicted combination of mental health conditions to understand the specific nuances in the discussion of bipolar disorder. Our results show that DeMHeM surpassed the baseline models and can be used to understand the multi-faceted discussion of mental health topics for a given community.**

***Index Terms*—datasets, neural networks, multitask learning, mental health, bipolar, depression, anxiety, Reddit, deep learning, topic model, keyphrase extraction**

I. Introduction

The stigma associated with many mental health conditions has made it very difficult for patients to share their successes and struggles freely. Bipolar disorder is one of the more common mental health conditions that are often subject to misunderstanding and discrimination. Social media and online forums such as Reddit have long been accessible avenues for a large section of the general population to share their experiences and seek support. Individuals suffering from such conditions will tend to prefer the anonymity of certain online spaces that also give them an opportunity to connect with like-minded people. One example of such a community is the "r/bipolar" subreddit, a Reddit subspace dedicated to those with bipolar disorder. From 2020 to 2022, this platform witnessed over 78000 posts of all types of content from all sorts of authors [4].

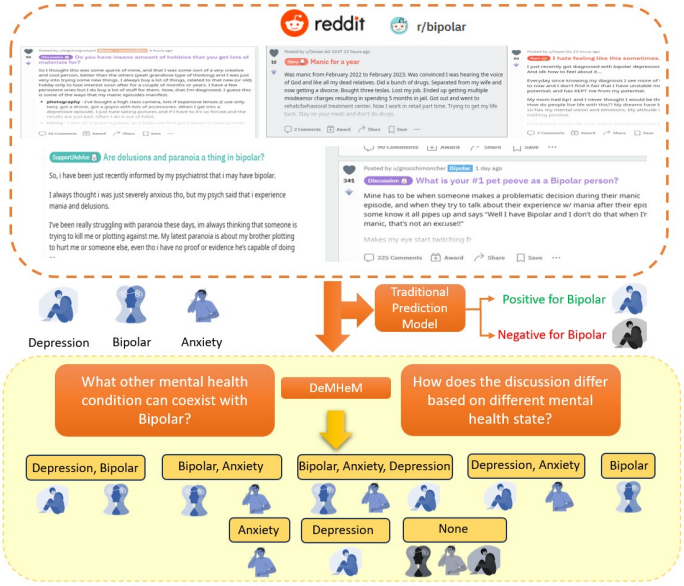

Fig. 1: In DeMHeM, the model predicts multiple conditions in a single post primarily for the analysis of a single condition, allowing for the distinction between specific states of the primary condition.

Previous research using similar online spaces tends to favor the more common and strongly signaled behaviors like suicidal thoughts and depression [11], [19], [34]. However, other more niche mental health disorders are not insignificant, yet unfortunately, they've been proven harder to classify due to their less

distinguishable indicators and smaller proportion of affected individuals [5]. Studies that do include rarer conditions tend to review a varied set of conditions, which is why we decided to focus on a single disorder while accounting for comorbidity in other behaviors. Unlike other works, our model aims to target the general population and predicts any given discussion's relation to the disorder, rather than the diagnosis status of a user as optimized by that user's explicit confirmation of diagnosis.

Research that aims to distinguish mental health conditions from each other often label every post in a dedicated discussion space by the topic of the space. However, this assumption is generally erroneous, yet a lot of literature studying Reddit data treats all posts in a bipolar subreddit as solely bipolar-centered [7], [16]. These studies may not provide further insight that can help mental health professionals to identify patterns of behavior and mitigate potentially grave situations like anxiety attacks, suicidal thoughts, or psychosis [2], [30]. Social media-based data mining and human-centered computing have always been fundamentally focused on utilizing the resources and advancement in the fields of machine learning and data mining to improve the quality of life for the whole community.

In this paper, we focus on bipolar disorder. Since people suffering from bipolar disorder can experience varying degrees of emotions ranging from hypomania or elation to extreme depression and anxiety, we believe that it is imperative to identify one or more of these conditions and analyze the nuances in the discussion pertaining to each. Specifically, we are focused on both depression and anxiety along with bipolar disorder to understand the context and nuances in these discussions.

To this effect, we propose a novel framework to predict the presence of the aforementioned mental health conditions from Reddit posts in the "r/bipolar" subreddit. While most of the existing works have tackled this issue by building predictive models for this kind of condition separately [10], [16], [28], we tackle the objective of modeling the coexistence and correlation among different mental health conditions by formulating it as a multitask learning (MTL) problem.

The existing works in mental health condition prediction have utilized lexical dictionaries and topic vectors as features for the prediction of depression and anxiety, and, in more recent years, pre-trained large language models have also been applied for predicting these different conditions. However, very few have targeted bipolar disorder and more specifically how it relates to anxiety and depression. This leads to a significant research gap in the literature along with the challenge of having no benchmark dataset available for the particular classification task. Hence, we collect and manually curate a dataset for bipolar, depression, and anxiety detection. We propose a novel multitask learning model **DeMHeM** (**De**scriptive **M**ental **He**alth predictions via **M**ulti-task learning) that redesigns the prediction problem as a multiple binary classification task and utilizes a multimodal feature extraction pipeline divided between shared and task-specific modules. Furthermore, we design a case study for posts pertaining to different categories through prediction-based clustering and keyphrase extraction.

The main contributions of this work can be outlined as follows:

- **Development of a novel multitask learning framework for mental health predictions.** DeMHeM accounts for the possibility of condition coexistence, or comorbidity, in a single submission by using multitask learning via a soft parameter sharing implementation with the auxiliary tasks of anxiety and depression detection. With the addition of auxiliary prediction tasks, the model is able to make more holistic and informative inferences that contribute to better in-depth analyses. The advantage of integrating MTL for the study of a single condition is improvement in both prediction performance and learned information about a post.
- **Implementation of a novel and effective multitask optimization algorithm.** MTL commonly uses tasks that are related but not similar, however, the framework implements MTL where all tasks are variations of mental health prediction. To encourage local convergence, regularization decay is proposed. To address the unbalanced nature of real-world data, Focal Loss is included to emphasize differences in post signals.
- **Conducting extensive experiments to demonstrate the effectiveness of the proposed framework** Reddit submissions were collected and processed from specific subreddits and manually annotated for three classification tasks. The effectiveness of the proposed DeMHeM framework along with its unique loss function was evaluated in an ablation study using the aforementioned dataset, and the framework as a whole is compared to common baselines to demonstrate overall performance gains.
- **Demonstration of framework usage via a case study on Reddit posts.** The proposed model is applied to real-world Reddit data from 2020 to 2022 for the detection of bipolar-related discussion. The predictions are used to conduct a case study by keyphrase extraction-based analysis on each predicted subcategory of mental health topics, and the results are displayed and discussed. The insights drawn from the study are theorized to be useful for associated professionals and social workers to better understand the nuances in the casual discussion of mental health topics.

II. Related Work

Our work merges multiple areas in data mining that are commonly studied independently of each other. In this section, we review existing work in these areas under three topics: multitask learning, social media-based data mining, and research on specific mental health conditions. We discuss how naturally broad social media data is used to detect various specific mental health conditions and behaviors, such as depression, anxiety, or schizophrenia. During the contextualization of our study, we indicate the various aspects of these works that our study of bipolar disorder addresses or incorporates.

A. Multi-Task Learning

Multitask Learning is a framework of machine learning that trains a single parameter space to perform multiple tasks. Aside from gaining a more diversely capable model, multitask learning improves performance by allowing signals from separate yet related tasks to affect the same parameters, causing better generalization and data and parameter efficiency [27], [36]. NLP has not been excluded from the benefits of this framework, and many studies have employed this technique with great success [37]. MTL consists of multiple parts including training framework, loss calculation, and model architecture, so there has been a growing number of approaches to its utilization. One popular MTL scheme is hard parameter-sharing, where models use a shared feature extraction module that feeds into task-specific prediction layers [8], and it has been shown to improve upon single-task learning frameworks in the notoriously difficult task of predicting mental health conditions [5]. Lokala et al. [20] used attention and a previous work [28] used message passing to implement MTL and improved performance on mental health-related tasks. Mrini et al. [21] employed soft parameter sharing, a technique where parameters across different tasks are encouraged to be similar, and exceeded other baseline MTL models in medical question understanding. In this work, we undertake joint learning with a soft parameter-sharing scheme using two auxiliary tasks to better understand a specific mental health condition.

B. Social Media-based Data Mining

Social media is a place where users can express themselves and discuss current events, making it a popular source of textual data for all sorts of tasks. Lee et al. [17] developed a system that forecasts future influenza activity via CDC datasets, and by enhancing their data with social media data streams, they achieve more accurate predictions than before. Wu et al. [33] were even able to effectively forecast oil markets by using the vast amount of news headlines related to the industry. Shinde et al. [31] review works centered around the forecasting of the COVID-19 pandemic, which demonstrates the capability of using social media to benefit the global population by predicting the activity of large-scale public health events. While research on tasks like event detection and disease forecasting are common, mental health conditions directly affect around a fifth of adults in the US [24], making it an urgent and beneficial area to study. In the next section, we detail previous works that use social media to analyze public mental health.

C. Mental Health-related Tasks

Research using social media data to study mental health behaviors is bountiful [22], [29], [35], [38]. Twitter is a frequently used source of social media text, and it has been used for tasks like event detection, disease forecasting, and geographic identification [9], as well as mental health prediction. Almouzini et al. [3] detected depression signals using Twitter data, and by using a supervised Latent Dirichlet Allocation (LDA) model, Resnik et al. [26] reported the various topics that such Twitter users talk about. Abboute et al. [1] made advancements toward real applications with socio-economic impacts by mining Twitter to detect posts of a suicidal nature and include a web interface with the purpose of having psychiatrists consult such posts. The recent COVID-19 pandemic has also presented a unique case for studying the effects on the mental health of individuals. Wu et al. [34] studied the relationship between COVID-19 infection and the risk of depression in social media, focusing more on specific users rather than a particular post. Chen et al. [7] find insights on depression and anxiety in the COVID-19 pandemic through Reddit data. Reddit is another social media platform, however, unlike Twitter, posts must be made under a specific group called a subreddit. Using Reddit data thus allows researchers to pick the general categories of post content, enabling them to access concentrated data and construct specific datasets; for this reason, Reddit is a decently common source for NLP tasks [2], [16], [19].

These studies tend to focus on the subject of anxiety and/or depression or related topics like suicide. While these conditions are prevalent and have strong impacts on individuals and communities, a multitude of other mental health conditions exist that make up a significant proportion of individuals with mental health conditions [24]. Gkotis et al. [12] trained a CNN to detect signals from conditions including borderline personality disorder, schizophrenia, bipolar disorder, and autism spectrum disorder. However, the study neglects to accommodate for the possibility of comorbidity and performs multi-*class* classification where a post can only have one labeled mental health condition. Benton et al. [5] trained separate MTL models to classify four main mental health conditions using a variety of sets of auxiliary tasks and showed significant improvements in AUROC score from single-task methods and previous works. Some recent works [11], [28] combine multitask learning with mental health prediction and demonstrated another advantage of MTL by discovering additional information about mental health conditions through predictions from auxiliary tasks. However, these works do not expand upon the specific study of more idiosyncratic conditions such as bipolar disorder.

These works show the benefits of performing a multi-*label* task which can simultaneously help models specialize in detecting specific conditions while also generalizing parameters to find better performance. However, many studies that include the broader space of mental health behaviors tend to lump many of them together [6], [32]. To the best of our knowledge, most bipolar-specific studies focus on individual-level, self-reported diagnosis data and do not use MTL with other auxiliary tasks [15], [30]. Thus, in our current work, we tackle a population-level study using human-annotated multi-label data and use MTL to perform an in-depth study on the specific mental health condition of bipolar disorders on Reddit.

III. Methodology

In this section, we outline the explicit goal of the model before elaborating on the architecture of the chosen model

for our DeMHeM framework and defining the modified loss function for training.

A. Problem Formulation

In this section, we define a general multitask learning objective and the notations that will be used to formulate and explain the equations for our model. A multitask learning model is trained in a supervised learning framework that aims to model a set of t tasks $T = \{T_1, T_2, T_3, ...T_t\}$ that are associated with a set of datasets $D = \{D_1, D_2, D_3, ..., D_t\}$ where each task T_i corresponds to a dataset D_i that contains labels Y_i, and each task's objective is to calculate $\hat{Y_i}$ where $\hat{Y_i}$ is the predicted target variable for T_i. In our formulation of the multitask model, we have a set of submissions $S = \{S_1, S_2, S_3, ...S_n\}$ where each submission can have multiple positive labels corresponding to each task in T. Hence, our supervised learning setting needs to learn the coexistence of multiple conditions within the same text instance by utilizing the same data instances for training each task. This results in a slight deviation from the traditional framework since the task set $T = \{T_1, T_2, T_3, ...T_t\}$ is associated with a single dataset $D = S$ where each submission S_j has ground truth vector $y = \{y_1, y_2, y_3, ...y_t\}$. Here, y_i is the one-hot ground truth corresponding to task i for S_j. The objective is to learn the function $f(X)$ that models Y where X is the feature vector for S and Y is the ground truth vector for all the tasks. We achieve this by learning both a separate loss function corresponding to each task and a joint loss function along with a regularization term, as further discussed in III-D.

B. Topic Modeling

When training an NLP model on embedded text, the model is limited only to the semantic meaning of the input text during inference. To incorporate more global information, we exploit the power of unsupervised topic modeling to provide context for submissions, and before the training of the main model, perform topic modeling on the training data. Each text is then associated with a corresponding topic vector $X_{topic} = \{t_1, t_2, t_3, ..., t_n\}$ where t_i is the probability that the text belongs to topic i and n is the total number of topics found. While commonly used for topic discovery and document analysis, topic modeling also provides an encapsulated context for a document's place in the corpus. By combining the topic probabilities into a single vector, we extract topic-level embeddings and are able to provide more information about documents. Latent Dirichlet Allocation is usually a first pick for topic modeling [23], but we decided to use a standard, pre-trained [1] BERTopic [13] model instead. The incorporation of similarity-based embeddings allows for easy clustering via hierarchical methods. Using agglomerative clustering, we restrict the topic model to produce a fixed 20 topics, allowing for consistency across runs.

[1]https://huggingface.co/bert-base-uncased

C. Multitask Learning Model

Our proposed multitask learning model builds on the observations of previous studies regarding the representation power of both sentence-level and topic-level embeddings for learning a shared latent feature space as well as task-specific semantic feature space [28] by building a multitask learning scheme that uses regularization decay of task-specific modules while also incorporating Focal Loss to facilitate joint learning.

Shared Feature Module: The shared module serves the purpose of extracting and compressing a document's semantic features. We begin by encoding the documents using a pre-trained Sentence-BERT [25], a model proven to give high-quality semantic features. In our design, the SBERT encodes the posts into 384-dimension vectors, which are then fed into a padded convolutional layer with 64 filters of size 5. A basic average is then applied to the now document matrix to compress it back into a vector, after which dropout is applied. We found that performing compression in the shared space rather than the task-specific space tended to improve performance; thus, we include a single fully connected layer with 32 outputs nodes followed by activation and dropout. The transformations for the shared feature module can be summarized as follows:

$$Z = Avg(Conv(X)), Z \in \mathbb{R}^{N \times d} \tag{1}$$

where $d = 384$ and N = batch size, then finally

$$Z_{final} = f(Dropout((W^T \cdot Z + b))), W \in \mathbb{R}^{h \times d} \tag{2}$$

where $f = ReLU$ is the activation function and $h = 32$ is the final hidden layer dimension for this module.

Task-Specific Module: The objective of the task-specific module is to interpret latent features found by the shared module as well as the topic model. During evaluation, the topic model encodes the unseen text data into topic vectors. As topic vectors already contain high-level information, they are directed straight to the task-specific module to assist with predictions. Since different tasks will find different topics more important than others, we process the topic vector with a single fully-connected layer to encourage the model to select only task-relevant information. After concatenating the processed topic vector with the extracted features from the shared module, the resulting tensor is passed through a multi-layer perception classification head. The final output of a single module indicates the predicted probability the text is classified with the corresponding task. The equation for the task-specific module can be expressed as follows:

$$H_1 = Concat(Z_{final}, X_{topic}), X_{topic} \in \mathbb{R}^{N \times n} \tag{3}$$

where X_{topic} is the topic level representation of batch size N, $n = 20$ is the number of topics, and $H_1 \in \mathbb{R}^{N \times (n+h)}$

$$H_t = W_t^T \cdot H_{t_1} + b, W_t \in \mathbb{R}^{h' \times h} \tag{4}$$

where H_t is the hidden representation for task t where $H_t \in$

$\mathbb{R}^{N \times h'}$ and

$$Y_t = f(H_{t_2}) \tag{5}$$

where $f = ReLU$ is again the activation function and Y_t is the prediction for task t.

D. Computation of Overall Loss

In multi-label tasks, loss is often computed by the sum of binary cross-entropies between the predicted probability of a positive label and the ground truth positive (1) or negative (0) label of each class. In our study, we also utilized a per-class weight factor α_i that emphasizes the cross-entropy of some classes more than others. This weighted loss function forms the basis of our loss, and we build upon it by adding two extra terms.

Focal Loss: Originally used for classification in object detection, Focal Loss [18] extends cross entropy by including a $(1 - p_t)^\gamma$ term that addresses class imbalance by strongly punishing confident misclassifications (i.e. when $\hat{y} = 0.05$ and $y = 1.0$) and weakly discouraging uncertain classifications (i.e. when $\hat{y} = 0.6$). Although the class distributions in our training data were not significantly skewed, we reasoned that including Focal Loss would help our model better understand the subtly distinct nature of bipolar disorder among the other mental health conditions and, in conjunction with MTM, improve generalization. Huang et al. [14] show findings that indicate class-balanced Focal Loss provides no noticeable improvement, possibly because the purpose of Focal Loss is to treat all classes equally and only correct those that are misclassified. Thus, in our study, we included this task-impartial joint loss function with focusing parameter $\gamma = 2$ that acted on the unbiased combined distribution of all predicted class probabilities and weighted its importance in the overall loss with a factor β.

Multi-Task Regularization Decay: Weight decay was used, however, to implement regularization for soft parameter-sharing, we also included a regularization loss function weighted by λ. In this study, we further utilize the similarity between each mental health prediction task and decay λ such that the importance of multitask regularization diminishes exponentially with each training epoch. Let a represent the exponent base and t be the t^{th} epoch of training so that λ_t denotes the value of λ during epoch t. It then follows that $\lambda_t = \lambda_0 a^t$. The intuition behind this method is that in addition to using a shared latent feature space, all task modules are forced to interpret these features in a similar, and hopefully more generalized, way for the earlier part of training, with the goal of better accuracy and faster convergence. Since the tasks are still unique, we use an exponential function to relieve the effects of this regularization term for the greater part of training, freeing the modules to converge to local minimums separately.

Let i represent the i^{th} task, n denotes the total number of tasks, and $\theta = \{\theta_1, \theta_2, ..., \theta_n\}$ denote the parameters of task i's specific module. Our overall loss function can then be computed as follows.

$$L_{total} = \sum_{i=1}^{n} [\alpha_i BCE(\hat{y}_i, y_i)] + \beta FL(\hat{y}, y) + \lambda L_{reg}(\theta) \tag{6}$$

Here L_{reg} calculates the sum of the L^2 norms of the difference (or Euclidean distance) between parameters of all possible pairs of task-specific modules. In other words,

$$L_{reg}(\theta) = \sum_{\{i,j\} \in \binom{n}{2}} \sqrt{(\theta_i - \theta_j)^2} \tag{7}$$

. We then used the Adam optimizer on a 1-cycle schedule to update the weights of our multitask model until convergence.

IV. Experiments

To assess the performance of the proposed DeMHeM, we design an extensive experimental setting to validate its effectiveness. These experiments aim to answer the following research questions:

- **RQ1:** How does DeMHeM perform compared to other baseline methods in our multitask scenario of bipolar, anxiety, and depression detection?
- **RQ2:** Does the soft/hard parameter sharing scheme or the incorporation of Focal Loss affect the model performance?
- **RQ3:** How does the imbalanced nature of the bipolar dataset affect the performance of the model compared to the baselines?
- **RQ4:** Does involving topic embedding for task-specific modules improve the performance?

A. Datasets

Reddit is a popular social media platform where users can start discussions by posting a submission of any length to a specific group called a subreddit. Denoted by an 'r/' prefix, subreddits center around a large variety of topics including mental disorders, making them a common source for mental health-related language data and thus form the source of our language data as well.

3800 randomly selected submissions were taken from "r/bipolar" in the period of January 1st, 2020 to December 31st, 2021. Each post was given three binary labels corresponding to their relevance to the class labels of bipolar disorder, anxiety, and depression. Control data was taken from "r/AskReddit", "r/Jokes", "r/CasualConversation", "r/CozyPlaces", etc. under the assumption that all submissions in these subreddits did not contain information related to the relevant classes. The data set constitutes of submissions from the various subreddits, balanced by class label counts. All of the data was extracted using pushshift.io, with the exceptions of submissions from r/AskReddit and r/Jokes, which were pulled from the one-million-reddit-questions and one-million-reddit-jokes data sets in Hugging Face's SocialGrep [2] data sets.

[2]https://hugging*-face.co/SocialGrep

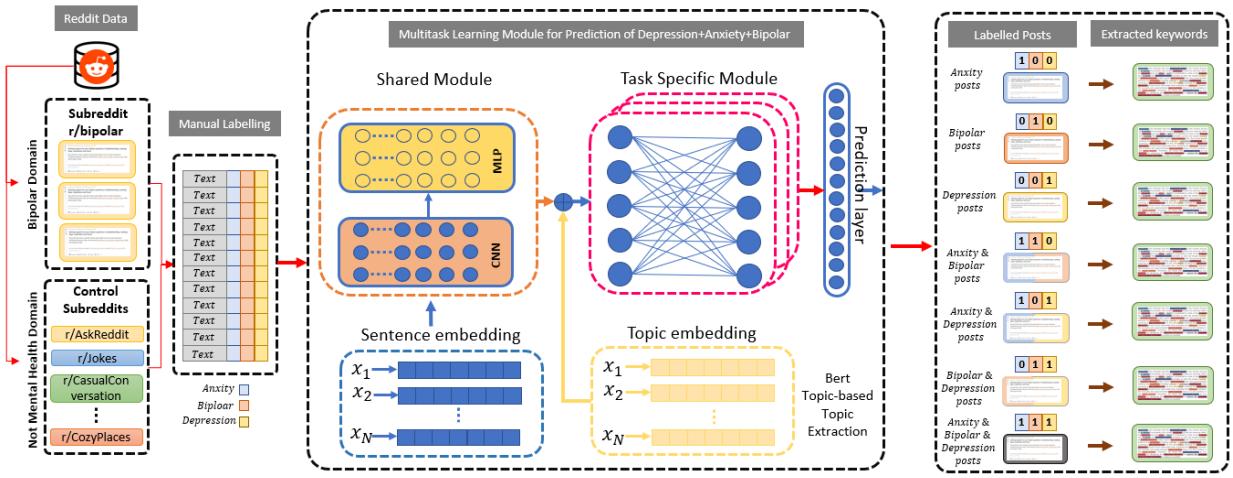

Fig. 2: The illustrative architecture of the proposed G2F framework.

Preprocessing: The conventional SBERT pipeline, which is also the embedding method we used, utilizes WordPiece tokenization, which is dependent on whitespace as well as punctuation. Thus, the preparation of our data only includes the removal of specific words. Aside from English stopwords, we also removed Reddit-specific artifacts such as "[removed]" and "[nan]", which indicate removed content or non-text content like images, or "", which are a result of untranslatable characters.

Labeling: Two annotators were assigned to label all 3800 Reddit posts, following the same guidelines. Our goal during labeling was not to diagnose users. Rather, it was to describe the relevance of a post to a specific condition. For a positive anxiety label, the post's author needed to either explicitly state they or someone else has/had anxiety or include anxiety-like language (e.g. words that indicate fear or worry) in their post. The depression label has similar guidelines, with some changes. A single strong key term like "suicidal" or "empty" is enough to merit a positive label, however, some explicit mentions of depression like those that neutrally refer to a bipolar depressive episode or posts that positively talk about depression did not get a positive depression label. The bipolar label was much more freely given out; a post only had to mention any bipolar keyword to get a positive label. A post only received a negative label if it was indeterminable that the post came from a bipolar-related subreddit (i.e. "I got diagnosed today" is related to the author's bipolar experience but the content is not specific to bipolar).

B. Experiment settings & Baselines

Baselines: As our work includes the combination of multiple different methods to improve performance, we gauge improvement by comparing DeMHeM to other common classifiers. Since these methods are not inherently multitask, unless otherwise stated, each baseline algorithm includes n separate models that perform binary classification for each task.

- **Logistic Regression (LR):** A common baseline model, Logistic Regression is a simple yet performant regression model that calculates the probability for a certain outcome given an n-dimensional input.
- **Random Forest (RF):** Random Forests utilizes wisdom in numbers by employing multiple decision trees to vote for a certain outcome.
- **k-Nearest Neighbors (kNN):** The k-NN algorithm assumes closeness in n-dimensional space correlates to semantic relatedness and classifies samples by using which class was most present among the k nearest neighbors of that sample.
- **Support Vector Machine (SVM):** SVMs are a powerful type of classifier that uses support vectors to find decision boundaries that differentiate between separate classes.
- **Multi-Layer Perceptrons:** MLPs are a dense, fully-connected network with activations between layers. Outputs are calculated via a sigmoid activation on the last layer to find the predicted probability of a positive label. In this study, we compare our model against two separate designs of MLPs.
  - **Merged (mMLP) -** A single MLP is designed with n output nodes corresponding to each task. This model serves to demonstrate the need for task-specific learning to accommodate the different tasks.
  - **Separate (sMLP)-** n separate MLPs are trained on each task, where no MLP can access information about other tasks or their models. This architecture serves to show task-specific adaptation while also indicating potential improvement via multitask learning.

C. Implementations Details

Our implementation of DeMHeM is written using PyTorch with CUDA support and 16 GB RAM. Since our model operates on other pre-trained models, the model's parameter count was low and we were able to load all data points into a single batch for training. Model parameters are uniformly initialized between -1 and 1, and embeddings are scaled via a

TABLE I: Overall performance of baseline methods in comparison to our method on a Balanced Dataset.

Model	Accuracy	Macro-F1	Micro-F1
LR	0.8189	0.7979	0.8171
kNN	0.7842	0.7683	0.7862
RF	0.7938	0.7549	0.7846
SVM	0.8201	0.7979	0.818
mMLP	0.8237	0.805	0.8224
sMLP	0.8201	0.8041	0.8206
DeMHeM	**0.8616**	**0.8457**	**0.8613**

TABLE II: Ablation Study with model variations controlling for inclusion of topic vectors (T), per-task weighting (W), Focal Loss (F), and regularization decay (RD)

T	W	F	RD	Accuracy	Macro-F1	Micro-F1
				0.8338	0.8099	0.8323
x				0.8410	0.8190	0.8398
x	x			0.8443	0.8222	0.8431
x	x	x		0.8504	0.8308	0.8500
x	x	x	x	0.8616	0.8457	0.8613
Hard Sharing				0.8544	0.8363	0.8542

standard scaler fitted on the training data. Labels are uniformly smoothed with $\epsilon = 0.05$ such that a ground truth vector like $\{0, 1, 0\}$ becomes $\{0.05, 0.95, 0.05\}$ during training. Results are obtained from optimized hyperparameters in 5-fold cross-validation. Specific scores are calculated by averaging the average scores for all binary classification tasks for all folds.

D. Results and Discussion

Overall Performance: To answer **RQ1**, we balanced the training data by raising the proportion of positively labeled "Anxiety" data to nearly 25% compared to the original 12% presently seen in the annotated dataset as well as removing many bipolar-only submissions, among other things. Table 1 details the performance of the baseline models on this data described in IV-B compared to our DeMHeM framework. The DeMHeM framework comes out on top with other models performing expectedly. However, we made an interesting observation. We hypothesized the mMLP model would perform worse than the sMLP model since the merged format should be less capable of finding unique features, yet it seems like the merged format is actually exemplary of a hard-sharing multitask model with a single task-specific node, since it actually improved generalization, potentially since the model had to learn a general interpretation of features with the last node making unique predictions.

Since the experiments were conducted on the dataset collected from the "r/bipolar" subreddit, we naturally had skewed training data since most posts would be bipolar-positive and fewer would be depression or anxiety-positive. This gives us a perfect opportunity to explore **RQ3**. Table 2 shows the accuracy and F-1 score of the model compared to baselines which shows a considerable improvement. All models take a sizeable hit to F1 scores in this set of data, however, Random Forest and Support Vector Machine fare considerably worse than before, possibly due to their preference for similarly-sized classes and smaller datasets. Although the proposed framework performs better on the balanced dataset, its performance on the unbalanced dataset is also relatively impressive. This could be due to additional regularization provided by the multitask regularization, as well as the well-suited Focal Loss. This observation is particularly encouraging given the ratio of the unbalanced data is more representative of the subreddit that we are basing the study on.

Ablation Study: To understand how different parts of the multitask learning model work and answer **RQ2** and **RQ4**, we design an ablation study to analyze the effectiveness of each method. As shown in Table 3, we design our ablation study by incrementally including different components of our DeMHeM framework. These are topic vectors, per-task loss weighting, Focal Loss, and multitask regularization decay. The last row details a hard parameter-sharing scheme that includes all the previous components except multitask regularization decay (as it is specific to soft sharing). All settings' hyperparameters are independently optimized such that our results detail the best performance we were able to achieve. As seen from the increasing Macro-F1 score, each addition does show improvement. Topic vectors give a considerable 0.0091 bump in Macro-F1, proving our hypothesis that global context improves local predictions. Task weighting had less of an improvement, which is likely because the soft sharing scheme is technically three separate models, so changing task weighting only changes gradient steepness for each task module. However, we reason that the improvement is due to the fact that the shallower gradients of some modules prevented the shared parameter space from favoring more prevalent classes when trained with soft sharing regularization. Although our dataset was relatively balanced, Focal Loss still proved helpful. Since most posts are bipolar-related, and depression-positive posts tended to have strong signals, these labels tend to be easier to classify than anxiety. We reason that Focal Loss helps likely because it assists in distinguishing the comparably weak signals from the anxiety task. The decay of the regularization between shared modules gave the largest improvement, proving our hypothesis that multitask regularization is less important after parameters settle in a collective minimum. In the earlier stages of our work, we defaulted to the hard-sharing scheme due to its simplicity and parameter efficiency. However, our results show that the soft-sharing scheme holds a slight edge over hard-sharing, possibly due to allowing more freedom for modules to extract and select task-necessary features.

V. Case Study

For our case study, we apply our model to predict the presence of bipolar disorder, depression, and anxiety for all 78k posts in the "r/bipolar" subreddit during the timeframe of the beginning of 2020 to the end of 2021. All posts

TABLE III: Overall performance of baseline methods in comparison to our method on an Unbalanced Dataset.

Model	Accuracy	Macro-F1	Micro-F1
LR	0.8494	0.7617	0.8421
KNN	0.8202	0.7300	0.8167
RF	0.826	0.6471	0.7953
SVM	0.8511	0.7403	0.8378
mMLP	0.8471	0.7638	0.843
sMLP	0.8517	0.7627	0.844
DeMHeM	**0.8806**	**0.8003**	**0.8768**

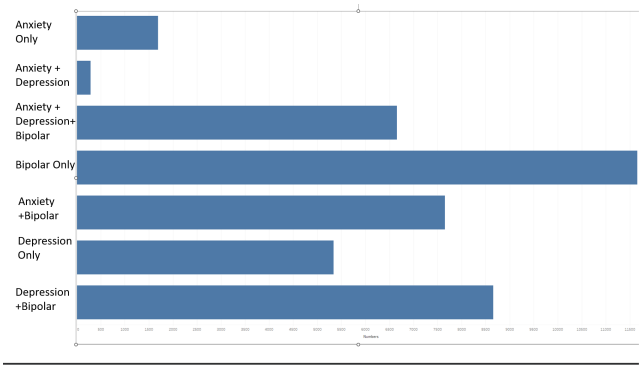

Fig. 3: Number of posts predicted to belong in each cluster

were clustered into 8 separate categories defined by their permutation of the three labels (e.g. all negative, just bipolar, bipolar and depression, etc.). Keyphrase extraction was then performed on each of these clusters to discover the topics that the users of the subreddit converse about.

Keyphrase Extraction: We considered various keyword/keyphrase extraction techniques such as YAKE, and TextRank, but ultimately settled on SingleRank. For each document in a cluster, the top keywords and their scores were calculated and added to the cluster's cumulative keyphrase map, such that when multiple documents have the same keyphrase, the scores for that keyphrase are added together. Listing the top keyphrases for each cluster now, however, would not be meaningful or interesting since many keyphrases such as "bipolar disorder" or "good" are generic and common to the whole subreddit. To locate the cluster-niche keyphrases, we excluded all keyphrases that appear in the top 50 keywords of more than $p = 0.5$ of clusters, where p is a proportion threshold that specifies the required uniqueness of a keyphrase. Figure 4 describes some of the top keyphrases extracted for each cluster using this method.

Finally, we apply our model to analyze the submissions in the "r/bipolar" subreddit to answer the following questions:

- **Q1: What is the frequency of mentions of other mental health conditions along with bipolar disorder for this specific subreddit?**
- **Q2: Among the group detected for different mental health categories, what are the top keywords and what insight can be drawn from that?**

A. Observations

To answer the research questions presented above the following important observations regarding the discussion of mental health in the bipolar community of Reddit are discussed.

To answer Q1: 78000+ posts were studied in "r/bipolar" between January 2020 and December 2021 by applying our proposed DeMHeM framework to predict the outcome for each bipolar, depression, and anxiety detection task. Out of these, at least 42000 posts were talking about users' experience with some form of bipolar disorder which constitutes more a large majority of posts, which is understandable given the subreddit is a platform for this particular community of people that wants to share their experience or seek help. However, this also signifies our assumption that it is not necessarily accurate to clump all posts in this subreddit under this category. As shown in Figure 3, we also observe quite a significant proportion of the posts in this category to be associated with depression which can be explained by the clinically proven phenomena of depressive episodes that people with one of the bipolar disorders experience. Overall, 27% of posts were talking about depression. Furthermore, we identify quite a significant number of posts discussing the feeling and condition of anxiety as well. If we look at the number of posts at the intersection between the discussions of bipolar and anxiety, it is clear that a significant percentage of people in the subreddit experience some form of anxiety. Upon observing the raw numbers belonging to each category we delve further into the specificity of discussion in these different categories through the keyword extraction for each of the clusters that the data is divided among.

To answer Q2: The keyphrase-based analysis on all category-based clusters can be seen in Figure 4. An interesting fact to reflect upon is that the keywords associated with any cluster involving a positive label for "Bipolar" has a mention of mania which is understandable given the fact that it is used to describe one of the frequent occurrences of mood shifts in patients suffering from bipolar disorder. However, a more interesting observation is that there are significant mentions of specific medications like "Lamictal" or "Gabapentin" that appear in the "Bipolar + Anxiety" category, which is different from the mention of medication like "Lithium" in the "Depression + Bipolar" category. The presence of category-specific keyphrases indicating a prescribed drug is a consistent observation as can be seen in the table. Furthermore, for posts only discussing "Depression", a sense of loneliness, guilt, and isolation is quite common. This can be observed through identified keywords such as "lonely", "hopeless" or "guilt". A minority of posts belonging exclusively to the "Anxiety" category more frequently mention social anxiety and paranoia. On the other hand, the keywords corresponding to the neutral category are often related to art and music. This demonstrates the nature of discussion in these subreddits is not exclusively limited to mental health issues but can often indicate how a

Category	Keywords
None	'artwork', 'med', 'Mondays mega thread', 'song', 'relatable',' care Sunday post self', 'post', 'rant','
Depression	'overwhelming pain', 'ptsd related', 'meds , 'lonely', 'hopeless', 'song', 'stable commitment tracking', 'vessel moods', 'deep depression', 'guilt', 'miserable'
Bipolar	'bipolar rage', 'bipolar psychosis', 'weed', 'relationship', 'bipolarity', 'trauma', 'mania psychosis'
Depression+Bipolar	'depression mania', depression bipolar', 'possible app bipolar disorder', 'best apps sites', 'magnifying feelings positive negative', 'lithium depression', 'chronic', 'severe mental disorders', 'mood chart seroquel'
Anxiety	'anxiety', vraylar' irreparable damage friendship', 'anxiety attacks', 'psychological test" , 'social anxiety', 'panic attacks', 'scared outcome', 'many people', 'paranoia'
Anxiety+Depression	'hospital desperate choice', 'end last session counselor', 'bad decision brewing', 'ashamed post', 'frustrated basic social issues', 'agitated detached body', 'wrong body', 'academic year', 'grad school', 'manic episode', 'support system mom boyfriend'
Anxiety+Bipolar	'lamictal', 'gabapentin' 'extreme anxiety', 'social anxiety', 'mania anxiety','manic abilify', 'anxiety attacks', 'unable manage',
All	'manic episode', 'last early summer', 'new job next week', 'way depressive manic episodes', 'mixed mania', 'app personal experience', 'substance abuse', 'freaking tired', 'illness', 'biggest problems',' impulsiveness',' extreme','adhd criterion similar bipolar criteria',

Fig. 4: Top keywords for each category

lot of people in this community like to share their creativity, hobbies, and other interests through candid self-expression.

VI. Conclusion

In this paper, we proposed a novel multi-task learning framework, DeMHeM, that utilizes the principles of inter-task parameter sharing, Focal Loss, and regularization decay to accurately model multiple categories of mental health conditions. By primarily focusing on the "r/bipolar" subreddit, we can learn the correlation between the discussion of anxiety, depression, and bipolar disorder. The annotated dataset produced through this process can be of significant importance when it comes to future research on this topic. Furthermore, the improved performance of the proposed model while tackling the inherent class imbalance problem common to a lot of mental health condition-related tasks indicates the validity of the design choices for the architecture of the model. This paper extends the work in social media-based prediction of mental health discussion by focusing on a very specific but often ignored topic of "bipolar disorder" with the intention of gaining significant insights about the nature of the discussion being had by people suffering from it. In order to understand how it is affected by their perceived experiences with mania, medication, therapy, familial support, or other underlying mental health issues like depression, suicidal ideation, or anxiety, we substantiate our findings through extensive experiments. The case study that applies DeMHeM and uses keyword extraction techniques on each category of predicted data gives quite a few significant insights into the aforementioned issues. In the future, one could aim for a more fine-grained annotation pertaining to "bipolar disorder" on the same data with the help of domain experts. We theorize that our framework can be applied to other conditions such as schizophrenia or ADHD, with the addition of other auxiliary prediction tasks other than depression and anxiety. In addition, if model-intrinsic explainability is also incorporated, it could be of immense use to social workers and mental health professionals and help address the various mental health conditions that affect our society today.

VII. Ethics Statement

In conducting this research, we have taken stringent measures to ensure the privacy and anonymity of individuals whose data has been analyzed. Given that the topic of mental health is sensitive and carries potential social stigmas, it was of paramount importance to us to handle the data with the utmost care. The Reddit posts used in this study are publicly available; however, all identifying information, including usernames and other potentially identifying markers, have been completely anonymized in both the dataset and any resulting publications. No individual Reddit users were contacted for the purpose of this research, and the dataset will not be used to identify any individual or their medical condition. Our aim is solely to gain insights into mental health conditions and their correlated nature for the benefit of scientific understanding and potential clinical application. Ethical review has been obtained to ensure that the methods and aims of this research are in line with standards for human subjects research. It is our hope that this work will contribute to the wider discourse on mental health, always with an eye toward respecting the privacy and dignity of individuals.

VIII. Acknowledgement

This research is supported in part by National Science Foundation grants CNS-2141095. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of any school board, NSF, or the U.S. Government.

References

[1] Amayas Abboute, Yasser Boudjeriou, Gilles Entringer, Jérôme Azé, Sandra Bringay, and Pascal Poncelet. Mining twitter for suicide prevention. In *Natural Language Processing and Information Systems*, pages 250–253. Springer International Publishing, 2014.

[2] Amanuel Alambo, Manas Gaur, Usha Lokala, Ugur Kursuncu, Krishnaprasad Thirunarayan, Amelie Gyrard, Amit Sheth, Randon S. Welton, and Jyotishman Pathak. Question answering for suicide risk assessment using reddit. In *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*. IEEE, January 2019.

[3] Salma Almouzini, Maher khemakhem, and Asem Alageel. Detecting arabic depressed users from twitter data. *Procedia Computer Science*, 163:257–265, 2019.

[4] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The pushshift reddit dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, 14:830–839, May 2020.

[5] Adrian Benton, Margaret Mitchell, and Dirk Hovy. Multi-task learning for mental health using social media text. *CoRR*, abs/1712.03538, 2017.

[6] Stevie Chancellor and Munmun De Choudhury. Methods in predictive techniques for mental health status on social media: a critical review. *npj Digital Medicine*, 3(1), March 2020.

[7] Justin Q Chen, Kevin Qi, Aaron Zhang, Mikhail Shalaginov, and Tingying Helen Zeng. COVID-19 impact on mental health analysis based on reddit comments. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, December 2022.

[8] Michael Crawshaw. Multi-task learning with deep neural networks: A survey. *CoRR*, abs/2009.09796, 2020.

[9] Oduwa Edo-Osagie, Beatriz De La Iglesia, Iain Lake, and Obaghe Edeghere. A scoping review of the use of twitter for public health research. *Computers in Biology and Medicine*, 122:103770, July 2020.

[10] Muskan Garg. Mental health analysis in social media posts: A survey. *Archives of Computational Methods in Engineering*, 30(3):1819–1842, January 2023.

[11] Soumitra Ghosh, Asif Ekbal, and Pushpak Bhattacharyya. A multitask framework to detect depression, sentiment and multi-label emotion from suicide notes. *Cognitive Computation*, pages 1–20, 2022.

[12] George Gkotsis, Anika Oellrich, Sumithra Velupillai, Maria Liakata, Tim J. P. Hubbard, Richard J. B. Dobson, and Rina Dutta. Characterisation of mental health conditions in social media using informed deep learning. *Scientific Reports*, 7(1), March 2017.

[13] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure, 2022.

[14] Yi Huang, Buse Giledereli, Abdullatif Köksal, Arzucan Özgür, and Elif Ozkirimli. Balancing methods for multi-label text classification with long-tailed class distribution. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 8153–8161. Association for Computational Linguistics, 2021.

[15] Glorianna Jagfeld, Fiona Lobban, Paul Rayson, and Steven H. Jones. Understanding who uses reddit: Profiling individuals with a self-reported bipolar disorder diagnosis. *CoRR*, abs/2104.11612, 2021.

[16] Zhengping Jiang, Sarah Ita Levitan, Jonathan Zomick, and Julia Hirschberg. Detection of mental health from reddit via deep contextualized representations. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*. Association for Computational Linguistics, 2020.

[17] Kathy Lee, Ankit Agrawal, and Alok Choudhary. Forecasting influenza levels using real-time social media streams. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 409–414, 2017.

[18] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *CoRR*, abs/1708.02002, 2017.

[19] Tingting Liu, Devansh Jain, Shivani R Rapole, Brenda Curtis, Johannes C. Eichstaedt, Lyle H. Ungar, and Sharath Chandra Guntuku. Detecting symptoms of depression on reddit. In *Proceedings of the 15th ACM Web Science Conference 2023*, WebSci '23, page 174–183, New York, NY, USA, 2023. Association for Computing Machinery.

[20] Usha Lokala, Aseem Srivastava, Triyasha Ghosh Dastidar, Tanmoy Chakraborty, Md Shad Akhtar, Maryam Panahiazar, and Amit Sheth. A computational approach to understand mental health from reddit: Knowledge-aware multitask learning framework. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1):640–650, May 2022.

[21] Khalil Mrini, Franck Dernoncourt, Seunghyun Yoon, Trung Bui, Walter Chang, Emilia Farcas, and Ndapa Nakashole. A gradually soft multi-task and data-augmented approach to medical question understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1505–1515, Online, August 2021. Association for Computational Linguistics.

[22] Usman Naseem, Adam G Dunn, Jinman Kim, and Matloob Khushi. Early identification of depression severity levels on reddit using ordinal classification. In *Proceedings of the ACM Web Conference 2022*, pages 2563–2572, 2022.

[23] Thin Nguyen, Bridianne O'Dea, Mark Larsen, Dinh Phung, Svetha Venkatesh, and Helen Christensen. Using linguistic and topic analysis to classify sub-groups of online depression communities. *Multimedia Tools and Applications*, 76(8):10653–10676, December 2015.

[24] National Survey on Drug Use and Health. 2021 nsduh annual national report. Nsduh annual report, National Survey on Drug Use and Health, 2023.

[25] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 2019.

[26] Philip Resnik, William Armstrong, Leonardo Claudino, Thang Nguyen, Viet-An Nguyen, and Jordan Boyd-Graber. 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*. Association for Computational Linguistics, 2015.

[27] Sebastian Ruder. An overview of multi-task learning in deep neural networks. *CoRR*, abs/1706.05098, 2017.

[28] Shailik Sarkar, Abdulaziz Alhamadani, Lulwah Alkulaib, and Chang-Tien Lu. Predicting depression and anxiety on reddit: a multi-task learning approach. In *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 427–435. IEEE, 2022.

[29] Ramit Sawhney, Harshit Joshi, Lucie Flek, and Rajiv Shah. Phase: Learning emotional phase-aware representations for suicide ideation detection on social media. In *Proceedings of the 16th conference of the European Chapter of the Association for Computational Linguistics: main volume*, pages 2415–2428, 2021.

[30] Ivan Sekulic, Matej Gjurković, and Jan Šnajder. Not just depressed: Bipolar disorder prediction on Reddit. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 72–78, Brussels, Belgium, October 2018. Association for Computational Linguistics.

[31] Gitanjali R. Shinde, Asmita B. Kalamkar, Parikshit N. Mahalle, Nilanjan Dey, Jyotismita Chaki, and Aboul Ella Hassanien. Forecasting models for coronavirus disease (COVID-19): A survey of the state-of-the-art. *SN Computer Science*, 1(4), June 2020.

[32] Ruba Skaik and Diana Inkpen. Using social media for mental health surveillance. *ACM Computing Surveys*, 53(6):1–31, December 2020.

[33] Binrong Wu, Lin Wang, Sirui Wang, and Yu-Rong Zeng. Forecasting the u.s. oil markets based on social media information during the COVID-19 pandemic. *Energy*, 226:120403, July 2021.

[34] Jiageng Wu, Xian Wu, Yining Hua, Shixu Lin, Yefeng Zheng, and Jie Yang. Exploring social media for early detection of depression in covid-19 patients. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 3968–3977, New York, NY, USA, 2023. Association for Computing Machinery.

[35] Kailai Yang, Tianlin Zhang, and Sophia Ananiadou. A mental state knowledge–aware and contrastive network for early stress and depression detection on social media. *Information Processing & Management*, 59(4):102961, 2022.

[36] Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5586–5609, December 2022.

[37] Zhihan Zhang, Wenhao Yu, Mengxia Yu, Zhichun Guo, and Meng Jiang. A survey of multi-task learning in natural language processing: Regarding task relatedness and training methods, 2023.

[38] Hamad Zogan, Imran Razzak, Shoaib Jameel, and Guandong Xu. Depressionnet: A novel summarization boosted deep framework for depression detection on social media. *arXiv preprint arXiv:2105.10878*, 2021.