Predicting Depression and Anxiety on Reddit: a Multi-task Learning Approach

Shailik Sarkar*, Abdulaziz Alhamadani*, Lulwah Alkulaib*†, and Chang-Tien Lu*
* Department of Computer Science, Virginia Tech, Falls Church, VA 22043 USA
‡ Department of Computer Science, Kuwait University, Kuwait
{shailik, hamdani, lalkulaib, ctlu}@vt.edu

***Abstract*—One of the strongest indicators of a mental health crisis is how people interact with each other or express themselves. Hence, social media is an ideal source to extract user-level information about the language used to express personal feelings. In the wake of the ever-increasing mental health crisis in the United States, it is imperative to analyze the general well-being of a population and investigate how their public social media posts can be used to detect different underlying mental health conditions. For that purpose, we propose a study that collects posts from "reddits" related to different mental health topics to detect the type of the post and the nature of the mental health issues that correlate to the post. The task of detecting mental health related issues indicates the mental health conditions connected to the posts. To achieve this, we develop a multi-task learning model that leverages, for each post, both the latent embedding space of words and topics for prediction with a message passing mechanism enabling the sharing of information for related tasks. We train the model through an active learning approach in order to tackle the lack of standardized fine-grained label data for this specific task.**

***Index Terms*—Topic, Multi-task, Neural Network, text classification, word embedding, active learning**

I. Introduction

"I just might kill myself" is one of the many suicidal expressions found on social media platforms that can be detected and prevented. Mental health issues, in general, have been one of the most critical issues in society. Several mental health conditions like depression, anxiety, suicidal ideation, or bipolar disorder. Suicide in the U.S. has been increasing in the past decades, becoming a national public health problem in the U.S. The number of deaths in the United States caused by suicide was 42,721 in 2018, and by 2019 the number of deaths by suicide was 47,467. Looking closer into(figure1) the rate of suicide per 100,000 in the U.S.A, we see an increase of 33% in the rate of suicide over 4 periods of time from 2005 to 2019. Suicide has become the 10th leading cause of death for adults in the U.S and the third leading cause of death among youth ages 10-14 and young adults ages 15-24. Suicidal ideation is often a product of underlying mental health conditions like depression, anxiety, or bipolar disorder taking adverse effects. Hence, the pacification of an individual with certain mental illness needs to account for these underlying conditions. United States has seen one of the worst cases of mental health crises in recent years, which has aggravated even further due to the outbreak of COVID-19. According to a KFF Health Tracking Poll run on July 2020, a few months after the lockdown enforcement in most states, there was a significant increase in sleeping or eating disorders, alcohol consumption, worsening of chronic conditions, substance abuse, etc. Furthermore, studies have shown that anxiety and other mental health issues increased significantly during the pandemic. Therefore, it is essential to identify and understand population-level mental health crises. With the ever-increasing accessibility, convenience, and the apparent protection of anonymity of social media platforms, more users tend to record their observations, discuss their

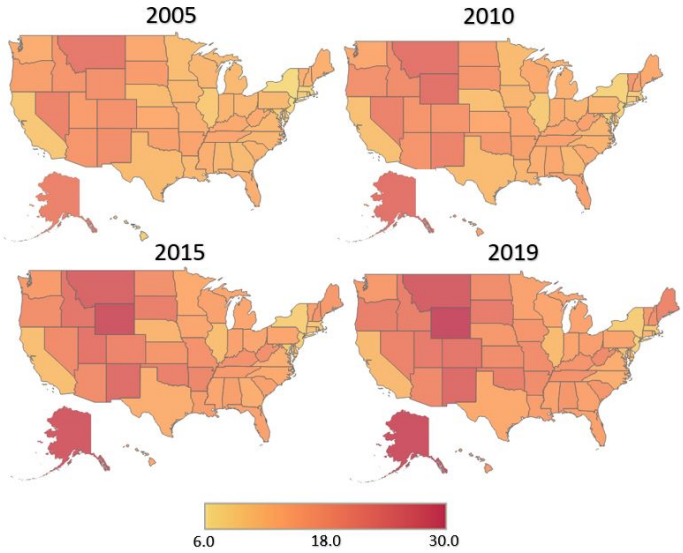

Fig. 1: A geographical heatmap of suicide mortality in different states of USA taken from CDC and social media and news media chatter about mental health crisis [1]

IEEE/ACM ASONAM 2022, November 10-13, 2022
978-1-6654-5661-6/22/$31.00 © 2022 IEEE

thoughts, and express their feelings candidly on their own mental health issues. Significant research has been done on applying natural language processing techniques for several tasks, including but not limited to sentiment analysis, event detection, depression or suicidal ideation detection, etc. Most of the existing work concentrates exclusively on identifying an individual user or post as indicating certain conditions. Most of the works use the subreddit topics as a target label. However, most of the time, a post in any of these topic-specific subreddits can be about multiple different mental health topics. Also, there needs to be a distinction between posts that are self-reporting as opposed to talking about someone else. For example, a subreddit like "r/bipolar" can have individuals with bipolar disorder or their family commenting on it. In contrast, a subreddit like "r/SuicideWatch" is mostly about individuals who are contemplating suicide or seeking help. These works mainly focus on using different linguistic dictionaries or topic modeling techniques as features for the prediction task while failing to address the shortcomings of some methods that do not consider a phrase as a separate entity. Furthermore, most works on this topic fail to address the relatedness among these different mental health conditions and the possibility of one post being related to multiple mental health conditions. Therefore, there is a significant challenge of creating a curated dataset that can indicate all these different nuances with proper labels so that a proper predictive model can be trained to identify these nuances in mental health discussion. Several works in the past have tackled this problem with techniques like dynamic query expansion based labeling process. However, recent development in active learning has also been proven beneficial in addressing the specific issue of curating an optimal framework for learning labels from an initially limited set of labeled data. As we are dealing with Reddit posts that are in hundreds of thousands, it makes for an ideal use case for the active learning paradigm of designing an optimal query to train a model iteratively. In relation to learning the nature of the post and the mental health categories, a Multi-task Learning(MTL) framework presents an ideal solution where predicting each label can be constructed as a separate binary classification task where the joint learning of parameters can address the relatedness of these different tasks which was not previously addressed by works in this field but has been applied in other NLP tasks like cyber threat detection and event detection. To address the challenge of predicting different mental-health conditions like depression and anxiety, we redesign the problem as a multi-task classification where one post can be detected as dealing with both depression and anxiety; or even neither, as often is the case in subreddits like "r/mentalhealth," "r/BPD" etc.

In this paper, we develop a novel framework for detecting different mental health topics in Reddit called Deep Active Multi-task Learning Model for Mental Health Topics(DeepAMTL-MH). DeepAMTL-MH uses an active learning framework based on a hybrid query generating technique which is a mixture of least confidence and highest entropy methods to iteratively train a multi-task learning encoder-decoder model to classify each post into a different category of mental health conditions. As our data sources, we use social media submissions from Reddit. We collect historical data from Reddit from each related subreddits as explained in detail in section IV. For extracting lexical features from our collected data, we experiment with different methods ranging from popular word embedding models(Glove, Word2Vec, BERT, etc.) to topic models like LDA and linguistic dictionaries indicating psychological aspects like LIWC(Linguistic Inquiry and Word Count). We enhance traditional topic modeling methods by considering both 1-gram and n-gram as separate tokens. The joint learning of the MTL model is based on a message-passing mechanism that leverages the correlation among different classification tasks. Finally, we geolocate the author of the posts to a specific state and analyze the difference in how the severity of different mental health conditions that can be inferred from these posts, and present a case study on how it was affected by the recent COVID-19 outbreak.

The main contributions of our work are summarized as follows:

- We tackle the lack of a fine-grained labeled dataset for Reddit that extends beyond topic specific subreddits by first curating a labeled dataset and then employing an active learning strategy to help with the training.
- We propose a novel multi-task learning model AMMNet that outperforms baseline models in the prediction of mental health conditions.
- We are the first to provide a model-level explanation behind our prediction due to the introduction of the task specific feature selector in our model. Our prediction ranks the most important topics associated with each prediction task.
- We show through extensive experiments that for domain-specific classification tasks such as this, a combination of document level embedding and topic distribution gives the best performance across all the tasks.

II. Related Work

In our work, multiple areas cross each other and sometimes are independent of each other. We first focus on existing work in Multi-task Learning (MTL) and active learning. For mental health, the outcome of human behavior on social media takes two directions. We discuss how social media data has been used to predict different health conditions. Then we describe works related to understanding many aspects of human mental behavior, such as depression, anxiety, suicidal thoughts, and more.

A. Multi-task and Active Learning:

Active learning has proven to be really useful when faced with data paucity issue especially in NLP tasks [8]. Furthermore, several active learning strategies like uncertainty based query, entropy based method or query by committee has proved to be particularly effective. [12], [24], [31] In

this paper we focus on employing one of these methods for training our final model. Multi-task Learning on the other hand has been a useful tool in NLP tasks like event detection [14]. There are several ways of designing a multitask learning models. One way to joint learn different objective is to use a shared feature extractor module followed by a task specific prediction layer and then by optimizing the combined loss function [32]. Some models use different task specific feature extractor while enabling message passing between different tasks through feedback loop or through intermediate node for storing information [17]. In our work we tackle joint learning by simultaneously learning both shared latent feature space and task specific feature map.

B. Social Media based health prediction

Social media has been a rich source of information for researchers interested in inferring different health conditions of the general population. While social media has previously been used in flu forecasting and other epidemiological applications, research on population-level prediction of different mental health outcomes is few. In this section, we will discuss works that use social media for mental health analysis of one or more individuals. [5] analyzed Reddit discourse on mental health conditions. They characterized self-disclosure and other related discussions. They also built a statistical model to understand the factors affecting social support for mental health. They also developed a language model to understand the social support for mental health issues. However, they do not focus on building a predictive model. [26] performs individual-level depression detection based on Twitter and a personal questionnaire, using an LDA topic model. [4] modeled language usage of individuals who have attempted suicide using social media data. Also, [3] used Twitter to build a language model for detecting several different kinds of mental health issues like ADHD. Similarly, [20] does a personality analysis based on Twitter on an individual level. Most of the works in this area are interested in individual-level identification of mental health conditions [6]. However, none of the abovementioned work focused on a spatial analysis of mental health outcomes. However, works like [9] go into the direction of spatial analysis, but they were exclusively focused on predicting heart disease occurrences. [23] makes a geospatial prediction of population well-being indicator based on Twitter discussion. More recently, [13] tried to estimate county-level well-being. While the well-being indicator is closely related to mental health, this work is not interested in quantitative analysis of mental health outcome statistics like suicide rate. Work on predicting suicide risk has been done on a much larger population level (e.g., for an entire country) [2], [25].

C. Mental health behavior

Large et al. [15] explored the range of preventative measures that can be a crucial catalyst in suicide prevention. The methods of prevention were categorized into 3 and each category-specific or general group of society; "universal" helps the whole population, "selective" helps high-risk groups, and "indicated" helps individuals. The work concluded that suicide categorizations result in a high false-positive rate and results in inaccuracies which may not be helpful for suicide prediction. This study [18] is a statistical research conducted on college students. The study aimed to develop a method to identify the significant forecasters of suicidal thoughts. The paper used random forest models with 70 potential forecasters, including social, sociodemographic, and substance abuse. We benefited from this study choose some of the keywords predictors. Continuing our survey to one of the state-of-art models to evaluate users on social media for suicidal risks. The study [21] formulates the problem as an ordinal regression problem and ranks users based on their risk of suicide. The reason for solving the problem as an ordinal regression is that it solves the situation when suicide risk evaluation happens, as not all wrong risk levels are equally wrong. The work presented a dual attention hierarchical model called SISMO, and they applied it to Reddit. They also added a human-in-the-loop to assist with interpretability. Our model learns significant points from this work, and we extend with our contribution, modifications, and enhancements. We address the problem of differentiating multiple mental health conditions within a single subreddit. Most of the existing works use the subreddit as the ground truth; however, it becomes tricky as subreddits about general mental health topics can vary from topics of suicide and depression to bullying or borderline personality disorder. This essentially creates a need for a curated multilabel dataset of Reddit posts where each post can have multiple labels, and each subreddit can have multiple labels that may even not be subreddit specific.

III. Methodology

For the task of mental health state prediction using Reddit submission text, we primarily address two challenges. First, the lack of reliable manually curated large datasets means that much time is needed to label data manually. We tackle this problem by building a hybrid active learning framework for the optimal labeling of the unlabeled dataset. Second, the correlation between different mental health conditions makes it necessary to understand the shared latent feature space and the task-specific semantic space for textual data. This challenge is addressed by reconceptualizing the multi-label classification task as a multi-task Learning(MTL) problem. Furthermore, we design different feature extraction techniques to capture the relevant features for shared and task-specific information.

A. Active Learning Module:

Active Learning strategies have proved to be particularly effective in reducing the labeling effort for curating an initially large training dataset.
We consider the following active learning strategies:

- Based on Least Confidence score: This queries instances for which the model generates maximum entropy, resulting in a set of data points for which the model is least certain. [16]

- Based on disagreement sampling: This strategy deviates from the uncertainty-based approaches by employing multiple classification models in training. After each training iteration, we use the vote and consensus probabilities over all the classifiers to calculate the disagreement for each instance. The query chooses the instances with maximum disagreement. To calculate the disagreement, we use the Kullback-Leibler divergence, which can be expressed as

$$D_{\mathrm{KL}}(P||Q) = \sum_i P(i) \ln \frac{P(i)}{Q(i)} \quad (1)$$

where Q is the consensus prediction of the set of learners and P is the prediction of individual learners.

Based on the query strategy, we take top n instances from the set of unseen data instances to update the training dataset for the multi-task Learning Classifier described in the following subsections.

B. N-gram Feature Extraction:

N-grams refer to a sequence of N words or characters. Our main intuition behind extracting the n-gram feature is to get a mixture of 1, 2, or 3-word sequences as tokens. For example

Let's consider the sentence: "I live in Washington DC."

A unigram model (n=1), stores this text in tokens of 1 word: ["I", "live", "in","Washington", "DC"]

A bigram model (n=2) stores this text in tokens of 2 words: ["I live", "live in", "in Washington", "Washington DC"]

In this scenario, the city "Washington DC" would not be recognized as an entity with the unigram since each token only stores one word. On the other hand, the bigram joins the words "Washington" and "DC" and allows the machine to recognize "Washington DC" as a single entity, thereby extracting the context from the text. That is why we hypothesize that N-gram tokens should be used as a precursor to our feature extraction pipeline. Each text sample can be represented as a collection of a topic. Topic Modeling is an unsupervised machine learning process that can represent the topic categories with which a body of texts can be associated. The topic model can divide the corpus into a fixed number of topics, each containing the most important words or phrases. Furthermore, it provides a token-level representation of a topic where each token is given a probability score for belonging to different topics. However, for the input of our model, we are interested in the probabilistic distribution of different topic categories for each text sample. In our work, we consider each submission a single text document. As shown in Figure2, on each of these documents, we use a topic model for extracting the topic level representation, which will be of the form $[p_1, p_2, p_i...p_n]$where $1 \leq$ i $\leq$n and p_i = probability score of topic i. We experiment with both LDA and BERTopic [11], [30].

C. Multitask-learning Module:

One of the advantages of using multi-task learning is the selection of both the common and the task-specific latent feature space. For textual data, we focus on two types of features: a) word embedding vector based on the embedding technique mentioned in subsection III-B and b) topic distribution vector based on the topic modeling technique mentioned in subsection III-B. As sentence-level embedding is more of a general feature representation for the text corpus, we pass this group of features through a shared component of interconnected hidden layers. We use a task-specific feature learning module for each post's topic-level representation.

Shared Feature Learning Module: This is a CNN-based model as presented in Fig2. The first layer of this architecture is an embedding layer. This represents the word-level embedding of each post with 200 dimensions. This should theoretically be better suited to capture the long-form nature of Reddit submissions. The word embedding used in our final design is a pretrained BERT-base model as it is better at capturing each word's contextual semantics. This layer is followed by the convolutional layer with the input of word embedding layers. This has 64 filters, each with a filter size of 5. Additionally, dropout is also applied. The output of this layer is passed through a max-pooling layer followed by a fully connected dense layer. This module works as a feature-sharing space for all the tasks. The output of this module is a tensor of length 32 which is then concatenated with the output vector of the Task-specific feature Learner module.

Task-specific Feature Learner: In the studies of text classification one important task is In the studies of text classification, one important task is to learn the task-specific features. For example, in posts related to anxiety terms like "Xanax," "stress," "fidgety," and "racing mind" will be exclusively present. At the same time, for the depression detection task, "die," "want to die," and "lonely" play a more significant role. We leverage unsupervised topic modeling techniques to capture this task-to-topic relationship. Hence, as input of this module for one data point we use a topic distribution vector of the form $\mathbf{T} = (\mathbf{t}_1, \mathbf{t}_2 \cdots \mathbf{t}_n)$ where t_i is the probability of the post belonging to the $i^{'}$th topic. This is a dense, fully-connected layer that takes as input the topic distribution vector $\mathbf{T} \in \mathbb{R}^{n \times d}$ where n is the number of instances in the input and d denotes the number of topics as discovered by the best performing topic model. For a fully connected dense layer of m number of nodes, the topic vector is multiplied by the weight vector of dimension $(d \times m)$ and the output of the layer can be written as follows:

$$\mathbf{H}_{feature} = \mathbf{XW} + bias \quad (2)$$

The output of this layer is thus a tensor of size $n \times p$. The main objective of having this layer as a buffer before concatenating it with the output from the shared feature extracting module is to penalize the weights associated with the unimportant topics. We learn task-specific topic importance by training two separate weight vectors for the universal topic distribution vector.

D. Group Lasso Penalty for feature sparsification

: To achieve feature sparsification, we propose using the group lasso regularization term. We are employing L_{12} regu-

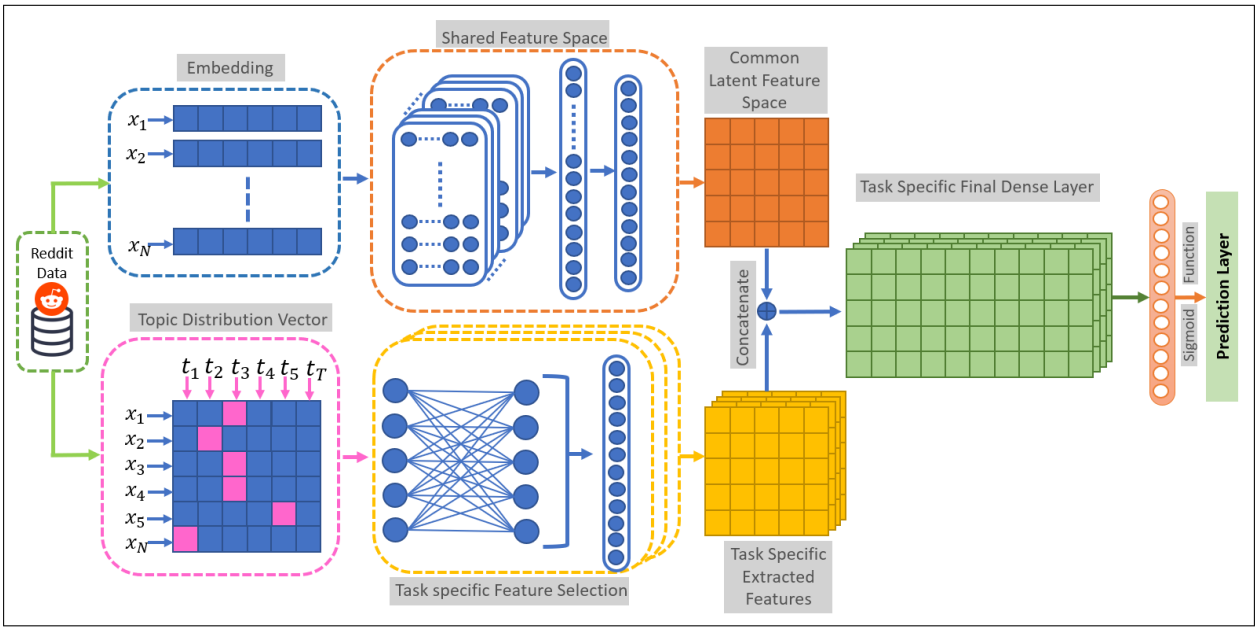

Fig. 2: The illustrative architecture of the proposed AMMNet method.

larization term written in the form of

$$\textbf{penalty} = \sum_{i=1}^{\mathbf{d}} \|W_i\|_2 \tag{3}$$

Here, W_i means each vector of the previously used tensor in equation 2 that is associated with a feature, where each element of the column vector indicates the weight associated with the corresponding hidden layer node. The intuition behind this is to mitigate the effect harmful features have on the original loss function as they increase the value of the function, making it hard to converge. Hence, by applying a group penalty at this stage to each feature, we minimize the value of all the weights connected to those unimportant features. Thus, we are potentially mitigating the effect of those harmful features while simultaneously providing an explanation as to which are the important ones, as depicted in subsequent section IV and table IV.

However, works on incorporating group lasso penalty in the neural network setting have also resulted in problems that non-differentiable penalty term poses when weights are very close to zero. Hence, to alleviate the problem, we incorporate the smoothing function for each vector as was previously adopted by [29] and other works [10], [28].

Concatenating groups of feature and Final Layer: This layer uses the concatenated vector from the shared feature extractor and task-specific feature selector as input. It is a fully connected dense layer followed by a sigmoid function for the classification output. It can be represented as:

$$\hat{Y} = \delta[relu(W \times Concat(F_{shared}, F_{topic}) + b)] \tag{4}$$

E. Computing Overall Loss Function for joint learning:

The overall loss function in each epoch includes computing both the binary cross-entropy loss and the group lasso penalty function for both tasks. Backpropagation is used for updating the weights at each layer based on the computed gradients. The overall loss function for the model will look something like this:

$$\mathbf{E}_{loss} = \sum_{j=1}^{\mathbf{n_{task}}} \alpha_j(\mathbf{E}_{BCELoss}(\hat{Y}, Y) + \beta \sum_{i=1}^{\mathbf{t}} \|W_i\|_2) \tag{5}$$

Here,n_{task} denotes the number of different tasks for the learner.α denotes a hyperparameter to control the weightage of the different task-specific loss function, and β is another hyperparameter designed to control the weightage of the group lasso loss function used for feature selection. Our objective is to minimize the loss while also updating the model's parameters until convergence.

F. Training Algorithm:

In this section, we explain the overall training process for the model. We divide the algorithm into two parts. First, we describe the disagreement-based active learning technique. Then in Algorithm 2, we describe the training for the multi-task learner, where the first stage is to use the embedding vector for the shared feature learning module and topic vector for the task-specific feature selector layer. The group lasso penalty term is applied to the weights of this layer. Next, the transformed feature vector is concatenated with the extracted feature vector from the shared module and passed through the final dense layers. At this point, we compute each task's overall loss function, including the penalty term. We minimize the

loss function by performing backpropagation on the weights of each layer according to its gradient and learning rate.

Algorithm 1 Active Learning strategy for MTL

Input: Set of m Learners $[L_1, L_2, L_3...L_m]$, Initial Labeled Training Set of n x =$[x_1, x_2, x_i...x_n]$ set of unseen data points x_{unseen}=$[x_{n+1}, x_{n+2}...x_s]$ number of instances to pick after each iteration = $n_{poolsize}$

while $i \neq iteration$ **do**
  $i = i + 1$
  P=[] /Probability Score
  while $j <= m$ **do**
    $model$ = $Train(x)$
    Y_{pred} = model.predict(x_{unseen})
    get probability score for each prediction and add them to P
  end while
  Q = average over all probability score for consensus
  Calculate $D_{\mathrm{KL}}(P||Q) = \sum_i P(i) \ln \frac{P(i)}{Q(i)}$ where i is a single instance of data
  Select top $n_{poolsize}$ from x_{unseen} and update x
end while

Algorithm 2 Multi-task Learning Model

Input: $X_{embedding} \in R^{N \times D1}$, $X_{Topic} \in R^{N \times T}$, $Y \in R^{N \times tasks}$;

Initialize parameters:W, Θ_1, Θ_2;

while t <= epoch **do**
  $H_{shared} = F_{\Theta_1}(X_{embedding})$
  while i <= tasks **do**
    $H_{topic} = X * W + bias$
    $H_{final} = Concat(\mathbf{H}_{topic}, \mathbf{H}_{shared})$
    $\hat{Y}_i = \mathbf{Dense}_{\Theta 2}(H_{final})$
  end while
  $E_{loss} = \sum_{j=1}^{n_{task}} \alpha_j (E_{BCELoss}(\hat{Y}, Y) + \beta \sum_{i=1}^{d} \|W_i\|_2)$
  Loss.backward()
  Update $W, \Theta 1, \Theta 2$
end while

.

IV. Experiments

This section will discuss our experimental setting, datasets, and results.

A. Datasets:

Any discussion about a topic on Reddit happens through different subreddits specific to a topic denoted with the prefix 'r/'. The submissions in these subreddits are what we are interested in. Submissions usually have a title and body and can be of variable length. We collect data from 'r/MentalHealth', 'r/SuicideWatch', 'r/Anxiety', 'r/bipolar' and 'r/BPD'. Earlier works used the subreddit topic as the label(E.g., Depression specific labels for depression subreddits, Anxiety Specific labels for anxiety subreddits). However, often we may find signs of a different mental health condition in other subreddits. For example, 'r/Anxiety' can have a post related to both anxiety and depression. The same holds for other subreddits. Hence, we focus on first creating a curated dataset of 65000 with three labels: 'Anxiety,' 'Depression,' and 'Others.' The collected data from Reddit was based on the most common depression and anxiety-related subreddits between the period of 2020-Jan to 2022-Jan. We employed PushShift.io [5] to extract all submissions from the previously mentioned subreddits. Once the dataset was collected, we compiled the data from the different subreddits into one dataset to prepare for fine-grained labeling of an initial small dataset of 6500 data points.

B. Experiment settings & Baselines

Both Topic Modeling and Word Embeddings are essential techniques to leverage for text classification. [19], [22], [27], [30] Therefore, we conduct our experiments across all base models using a combination of word embedding based features and topic modeling based features. For word embedding, we use pre-trained BERT model [7], and for topic modeling, we use primarily **LDA**and **BERTopic**. **LDA** is a well-known unsupervised topic modeling technique that has been used extensively in mental health research and has shown superior performance compared to other dictionary-based methods. **BERTopic** is an unsupervised topic modeling technique that uses BERT-based word embedding to form clusters based on UMAP and HDBSCAN that has been gaining popularity in recent times. We also focus our attention on statistical methods developed to represent textual data for large documents. To understand the significance of a token(word if 1-gram, phrase if 3-gram) in the context of a document and an entire corpus, we use **TF-IDF** features. The feature value for each token is calculated by calculating term and inverse document frequency. [1] We use 5000 of the 6500 data as Training Dataset. For all the base models while training, we split the training datasets on 80:20 training and validation split in a K-Fold cross-validation setting. The results in table I demonstrates the average performance across all metric using after 10 independently identically distributed runs. For our proposed Active Multi-task Learning MentalHealth condition prediction model(AMMNet), we train the model incrementally, starting from 2000 to 5000, with a pool of 300 data points at each iteration. TableIII shows the model's performance at each iteration on Test Set.

Baselines: As our survey has previously discussed, no other work tackled the specific problem of predicting depression, anxiety, or other(not belonging to these two conditions) categories of mental health conditions together as a multi-label classification task. Hence, we use traditional classification models in the form of both shallow and deep. The experimented models are as follows:

- **Support Vector Machine(SVM):** SVM is a powerful model for the classification task. It creates a decision boundary to differentiate between multiple classes.
- **Logistic Regression(Logistic)**:LR is another traditional classification model that has proved to be extremely effi-

TABLE I: Overall performance of baseline methods in comparison to our method on 5,000 Reddit submissions for Depression, Anxiety and Rest. Embedding(Emb), Percision (P), Recall (R), and micro-F1 (F1)

height$_{\text{Emb}}$	Logistic			Naive Bayes			KNN			SVM			Random Forest			MLP			AMMNet		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
TF-IDF	0.732	0.748	0.739	0.732	0.715	0.723	0.708	0.721	0.714	0.781	0.768	0.774	0.749	0.724	0.736	0.794	0.805	0.794	-	-	-
BERT	0.761	0.752	0.756	0.718	0.695	0.706	0.713	0.738	0.720	0.819	0.801	0.809	0.742	0.726	0.733	0.817	0.841	0.828	-	-	-
LDA	0.749	0.738	0.743	0.741	0.729	0.735	0.762	0.745	0.753	0.827	0.807	0.816	0.761	0.738	0.749	0.819	0.833	0.825	-	-	-
BERTopic	0.750	0.739	0.744	0.729	0.715	0.722	0.761	0.740	0.750	0.826	0.815	0.820	0.771	0.752	0.761	0.847	0.826	0.836	-	-	-
LDA+BERT	0.769	0.751	0.759	0.756	0.732	0.743	0.752	0.763	0.757	0.851	0.839	0.845	0.758	0.773	0.765	0.875	0.861	0.868	0.876	0.865	0.870
BERTopic+BERT	0.785	0.771	0.778	0.741	0.727	0.734	0.745	0.728	0.736	0.879	0.863	0.869	0.779	0.765	0.772	0.873	0.859	0.866	**0.881**	**0.867**	**0.874**

TABLE II: AMMNet fine-grained results

Category	**Percision**	**Recall**	**F1-Score**
Depression	0.898	0.879	0.888
Anxiety	0.865	0.853	0.859
Other	0.870	0.883	0.879

TABLE III: Active Learning training of AMMNet from initial labeled dataset of 2000

Training size	**Accuracy**
2000	0.832
2300	0.841
2600	0.839
2900	0.840
3200	0.842
3500	0.856
3800	0.869
4100	0.874
4400	0.871
4700	0.876
5000	0.875

TABLE IV: Most Important Topics for Each Task

Topic id	Category	Top Phrases/Words
23	Anxiety	”take” ”medication” ”doctor” ”day” ”meds” ”taking” ”panic_attacks
8	Depression	”help” ”really” ïve” ”depression” ”therapy” ”need” änyone” ”therapist” ”Struggling”
3	Depression	”cant” ”life” ”dont” änymore” ”dont_want” ”die” everything
14	Other	”work” ”job” ”home” ”go” ”day” ”covid
6	Anxiety	änxious” ”feeling” ”calm” öften” ”lot” ”älso” älways” ”worrying”

cient in any classification task. It models the probability of a discrete outcome given an input variable.

- **Naive Bayes**: This is a probabilistic classifier, based on 'Bayes Theorem' that is highly scalable, and has been extensively used for a variety of classification tasks.
- **K-nearest Neighbor(KNN)**: This is a non-parametric supervised classification model that relies on the feature vector distance estimation to predict the nearest data points to an instance in the same class.
- **Random Forest:** This is a decision tree based classification model which has been extensively used in text classification works. It is an ensemble learning method where the classification result depends on the class selected by most trees.
- **Multilayer Perceptron(MLP)**: MLP refers to a deep neural network model consisting of several fully-connected dense layer followed by an activation function.

In our experiments, we run our model AMMNet only on two specific combinations of feature vectors. Those are the only two experimental settings that facilitate the applicability of our framework due to the presence of two separate modules that deal with each set of features.

C. Results and Discussion:

Table I details results across different experimental settings using F-1, Recall, and Precision metrics. The models were all trained on the curated labeled dataset of 6500 data as explained in subsection IV-B. From the table, it can be seen that just the TF-IDF representation of textual data does not produce satisfactory results even in the case of MLP and SVM, which are two of the best performing among the baselines. However, using the BERTopic method to extract features significantly improves that performance. It is important to note that the topic modeling in BERTopic is based on TF-IDF scores across all the documents, even though the BERT embedding vector is used to calculate the distance in the embedding space. This supports our hypothesis that features extracted from unsupervised topic models can be highly effective for text classification tasks which is further strengthened by the performance of SVM and SVM+BERT experimental settings across all models. On the other hand, the BERT-based embedding of the document also supports our hypothesis of using BERT-based document level embedding as features for the classification task. This brings us to our observation regarding the primary hypothesis of redesigning the classification task as a multi-task learning problem where different mechanisms extract features from these two sets of features. It can be inferred from the table that AMMNet produces the best scores in all three metrics. However, SVM and MLP models do come close in terms of Precision.

In Table III we detail the accuracy scores of the proposed model on each iteration. From the initial observation, it is clear that the accuracy of the model improves steadily, more or less. This proves that even with a small dataset of a few thousand

text instances, a model can be trained with each iteration involving the training of an expanded dataset. This could potentially pose trouble for other computationally complex models, but that could be something addressed in future work. In Table II we see how the model performs individually for fine-grained prediction of depression and anxiety. The model clearly performs best for depression then compared to anxiety. This could be related to the fact that terms associated with depression are often associated more easily than anxiety-specific terms as it usually involves mentions of death and suicide as opposed to a more generalized sub-feature space for anxiety.

D. Most important Topics Discovery:

As our model enforces group lasso penalty term in the task-specific feature selector layer, it enables us to understand the most important features for each task. Furthermore, the corresponding Topic Modeling technique will give significant insight into these topics by listing the top keywords for each topic. As shown in TableIV the topic that is more telling for detection of depressive posts has words like "depression", "die" while for Anxiety some of the most important topics include phrase such as "panic attacks", "medication", "anxious". This can be a very useful tool while applying this model on a domain specific classification task.

V. Conclusion

In this paper, we proposed a novel Active Multi-task learning model AMMNet, that extracts task-specific features in the form of topics and learns from a shared feature space of document-level embedding. Our framework expands mental health prediction on Reddit from a subreddit-specific approach to a more general versatile input space. We successfully show that the combination of word embeddings and topic modeling can be leveraged for such domain-specific tasks. We tackle the data paucity issue in this domain by successfully adopting an active learning approach and expanding on the curated labeled dataset that will be available upon request. Due to the sensitive nature of the data, only after rigorous ethical screening should we make the data available. The paper also provided significant insight into the importance of different topics for predicting a given category of mental health conditions. We substantiate our findings through extensive experiments. A future direction of this work could be to look into specific mental disorders like "OCD" "BPD," or "bipolar" and try to predict or detect such conditions with explainability, as shown in this paper.

VI. 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] Akiko Aizawa. An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, 39(1):45–65, 2003.

[2] Marshall Burke, Felipe González, Patrick Baylis, Sam Heft-Neal, Ceren Baysan, Sanjay Basu, and Solomon Hsiang. Higher temperatures increase suicide rates in the united states and mexico. *Nature climate change*, 8(8):723–729, 2018.

[3] Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. From adhd to sad: Analyzing the language of mental health on twitter through self-reported diagnoses. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, pages 1–10, 2015.

[4] Glen Coppersmith, Ryan Leary, Eric Whyne, and Tony Wood. Quantifying suicidal ideation via language usage on social media. In *Joint Statistics Meetings Proceedings, Statistical Computing Section, JSM*, volume 110, 2015.

[5] Munmun De Choudhury and Sushovan De. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Eighth international AAAI conference on weblogs and social media*, 2014.

[6] Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 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, 2016.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. pages 4171–4186, June 2019.

[8] Liat Ein Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. Active learning for bert: an empirical study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7949–7962, 2020.

[9] Johannes C Eichstaedt, Hansen Andrew Schwartz, Margaret L Kern, Gregory Park, Darwin R Labarthe, Raina M Merchant, Sneha Jha, Megha Agrawal, Lukasz A Dziurzynski, Maarten Sap, et al. Psychological language on twitter predicts county-level heart disease mortality. *Psychological science*, 26(2):159–169, 2015.

[10] Mauro Forti, Paolo Nistri, and Marc Quincampoix. Generalized neural network for nonsmooth nonlinear programming problems. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 51(9):1741–1754, 2004.

[11] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

[12] Alex Holub, Pietro Perona, and Michael C Burl. Entropy-based active learning for object recognition. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8. IEEE, 2008.

[13] Kokil Jaidka, Salvatore Giorgi, H Andrew Schwartz, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. Estimating geographic subjective well-being from twitter: A comparison of dictionary and data-driven language methods. *Proceedings of the National Academy of Sciences*, 117(19):10165–10171, 2020.

[14] Taoran Ji, Kaiqun Fu, Nathan Self, Chang-Tien Lu, and Naren Ramakrishnan. Multi-task learning for transit service disruption detection. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 634–641. IEEE, 2018.

[15] Matthew Michael Large. The role of prediction in suicide prevention. *Dialogues in clinical neuroscience*, 20(3):197, 2018.

[16] David D Lewis and William A Gale. A sequential algorithm for training text classifiers. In *SIGIR'94*, pages 3–12. Springer, 1994.

[17] Pengfei Liu, Jie Fu, Yue Dong, Xipeng Qiu, and Jackie Chi Kit Cheung. Learning multi-task communication with message passing for sequence learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4360–4367, 2019.

[18] Melissa Macalli, Marie Navarro, Massimiliano Orri, Marie Tournier, Rodolphe Thiébaut, Sylvana M Côté, and Christophe Tzourio. A machine learning approach for predicting suicidal thoughts and behaviours among college students. *Scientific reports*, 11(1):1–8, 2021.

[19] Yuanhan Mo, Georgios Kontonatsios, and Sophia Ananiadou. Supporting systematic reviews using lda-based document representations. *Systematic reviews*, 4(1):1–12, 2015.

[20] Daniel Preotiuc-Pietro, Jordan Carpenter, Salvatore Giorgi, and Lyle Ungar. Studying the dark triad of personality through twitter behavior. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, pages 761–770, 2016.

[21] Ramit Sawhney, Harshit Joshi, Saumya Gandhi, and Rajiv Ratn Shah. Towards ordinal suicide ideation detection on social media. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 22–30, 2021.

[22] Timo Schick and Hinrich Schütze. Bertram: Improved word embeddings have big impact on contextualized model performance. 2020.

[23] Hansen Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Richard E Lucas, Megha Agrawal, Gregory J Park, Shrinidhi K Lakshmikanth, Sneha Jha, Martin EP Seligman, et al. Characterizing geographic variation in well-being using tweets. In *Seventh International AAAI Conference on Weblogs and Social Media*, 2013.

[24] H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 287–294, 1992.

[25] Takanao Tanaka and Shohei Okamoto. Increase in suicide following an initial decline during the covid-19 pandemic in japan. *Nature human behaviour*, 5(2):229–238, 2021.

[26] Sho Tsugawa, Yusuke Kikuchi, Fumio Kishino, Kosuke Nakajima, Yuichi Itoh, and Hiroyuki Ohsaki. Recognizing depression from twitter activity. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 3187–3196, 2015.

[27] Junaid Abdul Wahid, Lei Shi, Yufei Gao, Bei Yang, Yongcai Tao, Lin Wei, and Shabir Hussain. Topic2features: a novel framework to classify noisy and sparse textual data using lda topic distributions. *PeerJ Computer Science*, 7:e677, 2021.

[28] Jian Wang, Chen Xu, Xifeng Yang, and Jacek M Zurada. A novel pruning algorithm for smoothing feedforward neural networks based on group lasso method. *IEEE transactions on neural networks and learning systems*, 29(5):2012–2024, 2017.

[29] Jian Wang, Huaqing Zhang, Junze Wang, Yifei Pu, and Nikhil R Pal. Feature selection using a neural network with group lasso regularization and controlled redundancy. *IEEE Transactions on Neural Networks and Learning Systems*, 32(3):1110–1123, 2020.

[30] Xing Wei and W Bruce Croft. Lda-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 178–185, 2006.

[31] Yi Yang, Zhigang Ma, Feiping Nie, Xiaojun Chang, and Alexander G Hauptmann. Multi-class active learning by uncertainty sampling with diversity maximization. *International Journal of Computer Vision*, 113(2):113–127, 2015.

[32] Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 2021.