Analyzing Prediction of Depression and Anxiety on Reddit: a Multi-task Learning Approach through GMMTL

Sarkar Sarkar

shailik@vt.edu

Virginia Tech

Abdulaziz Alhamadani

Virginia Tech

Sristhi Behal

Virginia Tech

Lulwah Alkulaib

Virginia Tech

Chang-Tien Lu

Virginia Tech

Research Article

Keywords: multitask-learning, active-learning, topic-modeling, mental-health, depression, anxiety

Posted Date: April 3rd, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2759059/v1

Additional Declarations: No competing interests reported.

Analyzing Prediction of Depression and Anxiety on Reddit: a Multi-task Learning Approach through GMMTL

Shailik Sarkar[1*], Abdulaziz Alhamadani[2*], Srishti Behal[1], Lulwah Alkulaib[1] and Chang-Tien Lu[1]

[1,2,3,5]Department of Computer Science, Virginia Tech, Street, Falls Church, 22043, VA, USA.
[4]Department of Computer Science, Kuwait University, Street, City, 10587, State, Kuwait.

*Corresponding author(s). E-mail(s): shailik@vt.edu; hamdani@vt.edu;
Contributing authors: behalsrishti08@gmail.com; lalkulaib@vt.edu; 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.However, it is also important to be able to detect unnecessary post that are directly talking about the topic of interest. To achieve this, we expand our 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. We further add control group in the dataset to detect non-mentalhealth related post. Finally, we conduct a case study to show what are the important features in different mental health condition predictions.

Keywords: multitask-learning, active-learning, topic-modeling, mental-health, depression, anxiety

1 Introduction

The prevalence of mental health conditions such as depression, anxiety, and bipolar disorder remains a significant concern in contemporary society, impacting a substantial number of individuals. One issue that has garnered considerable attention in recent years is the escalating rates of suicide in the United States, which has emerged as a critical public health problem. According to statistics, the number of suicides in the country has been on a steady upward trend, with 42,721 deaths recorded in 2018, rising to 47,467 in 2019. Analysis of the suicide rate per 100,000 people in the US (Figure 1) reveals a 33% increase in the suicide rate over four periods from 2005 to 2019. Sadly, suicide has become the 10th most common cause of death for adults and the third most common cause of death for young individuals between the ages of 10 and 24. Suicidal ideation often stems from underlying mental health issues, exacerbating their impact. Therefore, it is crucial to consider these underlying issues while addressing the mental health needs of individuals with such illnesses. The COVID-19 pandemic has further aggravated an already severe mental health crisis in the United States. A KFF Health Tracking Poll conducted in July 2020, a few months after the majority of states started implementing lockdown measures, revealed a significant increase in sleeping or eating disorders, alcohol consumption, the worsening of chronic conditions, substance abuse, and other related issues. Additionally, research has indicated a significant increase in anxiety and other mental health problems during the pandemic. As a result, it is essential to identify and address mental health emergencies at a population level.

Open discussions about mental health, particularly from individuals experiencing it firsthand, are not often highlighted in the media. However, in recent years, social media platforms like Reddit have witnessed a rise in such conversations. With the ease of access and the promise of anonymity, people are increasingly sharing their struggles with mental health openly and honestly. It is crucial to provide individuals with a safe space to seek assistance and discuss their experiences, which is mostly facilitated by subreddits. For instance, subreddits dedicated to addressing specific mental health conditions, such as bipolar disorder, have seen a significant surge in posts over the past few years. Unlike other data sources like Twitter, where users face restrictions on both

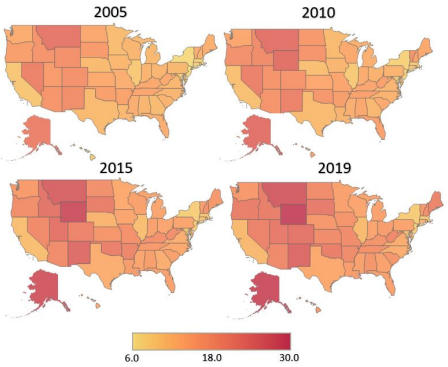

Fig. 1 A geographical heatmap of suicide mortality in different states of USA taken from CDC and social media and news media chatter about mental health crisis
[1]

brevity and privacy/anonymity, platforms without character constraints allow for more subtle and context-specific details to be shared.

Natural language processing techniques have been used for a long time to address the problem of identifying mental health conditions. Previous studies have used a combination of topic modeling, sentiment analysis, linguistic dictionaries, and deep learning models. However, these studies have often failed to establish benchmarks for acceptable features, data sets, or models that provide meaningful interpretation of important features.

Most related studies have used Twitter as the primary data source, which limits meaningful insights due to the brevity of post length and the lack of annotated datasets. However, in recent years, more studies have started to use Reddit as a data source. In a previous study, we used data from subreddits "r/Anxiety" and "r/Depression" to formulate a multi-task learning solution to the problem of mental health. Our proposed framework, Deep Active Multi-task Learning Model for Mental Health Topics (AMMNet) [1], used an active learning framework based on a hybrid query generation technique to curate a labeled dataset consisting of submissions related to depression and anxiety.

In this paper, we aim to answer the following research question:

- Can we differentiate between depression, anxiety, and non-mental health-related posts to enable mental health professionals to weed out unrelated posts from the vast array of social media data?
- How can we identify the most important topics corresponding to each task and is the explanation insightful?

To improve the accuracy of identifying mental health-related discussions, the model must distinguish not only between different mental health topics but also between mental health and non-mental health-related discussions. Therefore, our study uses both mental health-related subreddits and general subreddits as data sources. We consider submissions from unrelated subreddits as a control population for our labeled data on depression and anxiety. This

approach allows us to examine the correlation between different mental health condition prediction tasks. We propose a novel graph-based message passing mechanism in our multi-task learning model that enables the model to learn the joint probability.

Our pipeline utilizes lexical features extracted from various word embedding models (such as SBERT, BERT, and Word2Vec) and semantic features from topic modeling, linguistic dictionaries, and sentiment analysis models (such as LDA, BERTopic, LIWC, and EMPATH). We hypothesize that although each task-specific module's semantic feature learner learns different weights for their respective classification tasks, there is still a correlation among these tasks that cannot be addressed by jointly learning three separate loss functions. This can be solved by utilizing the proposed graph-based approach to message passing, which additionally learns the joint probability for all the task-specific objectives. Our "Graph-based Message-passing Multi-task learning model" (GMMTL) improves upon previous results in predicting mental health conditions and can differentiate non-mental health-related posts from those discussing depression or anxiety.

Moreover, we conduct a comprehensive case study using the explainable module which allows us to analyze most important features in posts for depression and anxiety or neither, as often seen in subreddits like "r/MentalHealth" and "r/BPD."

The main contributions of our work are summarized as follows:

- Our approach builds on our previous work to show how active learning can be used to solve the problem of manual labeling in issues as fine-grained as this one.
- Our proposed model, GMMTL, is a novel message passing multi-task learning approach that surpasses the performance of base models in predicting mental health conditions by introducing the concept joint probaility learning.
- We introduce an extensive experimental settings for a case study intoo the explainability of mental-health prediction, By ranking the most important topics associated with each prediction task, we are able to provide better insight into the reasoning behind our model's output, an improvement on our previous explainability module.

2 Related Work

Our work draws upon methodologies from multiple research areas that are often studied independently. We review the existing literature on multitask learning and active learning in the context of our proposed multitask learning model and the active learning framework we employed for initial dataset creation. We also examine previous approaches in data science for predicting various health conditions using social media. Finally, we narrow our focus to the specific task of mental health prediction using social media.

2.1 Multi-task and Active Learning:

Research has demonstrated that active learning is a powerful solution for addressing data scarcity issues in NLP tasks[2]. Various active learning strategies, including uncertainty-based query[3], entropy-based method[4], and query by committee[5], have been shown to be effective. In this study, we focus on utilizing one of these methods to train our final model.

Similarly, multi-task learning has also proven to be a valuable tool in NLP tasks, such as event detection[6]. There are multiple approaches to designing multi-task learning models. One approach to joint learning is to use a shared feature extractor module, followed by a task-specific prediction layer, and optimizing the combined loss function[7]. Other models utilize different task-specific feature extractors while allowing message passing between tasks through feedback loops or intermediate nodes for storing information[8]. In our previous work, we simultaneously learned both the shared latent feature space and task-specific feature map to tackle joint learning[7]. In this paper, we expand our methodology by introducing a graph-based message passing mechanism that assists in learning the joint probability for different tasks.

2.2 Social Media based health prediction

Social media data has been extensively used by researchers to infer various health conditions of the general population. Zhao et al. utilized social media data for flu forecasting[9], and Comito et al. showed that integrating social media data into COVID-19 forecasting models improved their effectiveness[10]. Recently, Jahanbin et al. used Twitter and web news as a data source for predicting monkeypox outbreaks[11]. While social media has been used in flu forecasting and other epidemiological applications, there are few studies on population-level prediction of different mental health outcomes. In the next section, we will discuss works that use social media for the mental health analysis of one or more individuals.

2.3 Mental health behavior

In recent years, social media has been extensively used for mental health research. Some works have analyzed social media discourse on mental health conditions and built statistical models to understand factors affecting social support for mental health. For instance, de Choudhury et al. [12] characterized self-disclosure and related discussions on mental health on Reddit, while Tsugawa et al. [13] performed individual-level depression detection using Twitter and a personal questionnaire. Coppersmith et al. [14, 15] modeled language usage on social media to detect suicide attempts and various mental health issues like ADHD, respectively. Similarly, Preoţiuc-Pietro et al. [16] performed personality analysis based on Twitter data on an individual level. However, most of the works in this area are focused on individual-level identification of mental health conditions [17].

Some works have gone in the direction of spatial analysis, but they have mostly focused on predicting heart disease occurrences [18] or well-being indicators [19, 20]. None of the mentioned works focused on a spatial analysis of mental health outcomes.

Large et al. [21] explored preventative measures for suicide prevention, categorizing them into universal, selective, and indicated methods. They concluded that categorizations result in a high false-positive rate, which may not be helpful for suicide prediction. On the other hand, MacAlli et al. [22] conducted statistical research on college students, aiming to develop a method to identify the significant forecasters of suicidal thoughts. They used random forest models with potential forecasters, including social, sociodemographic, and substance abuse.

Sawhney et al. [23] formulated the problem of evaluating users on social media for suicidal risks as an ordinal regression problem and ranked users based on their risk of suicide. They presented a dual attention hierarchical model called SISMO and applied it to Reddit, adding a human-in-the-loop to assist with interpretability. Our work extends from Sawhney et al.'s study by differentiating multiple mental health conditions within a single subreddit. This 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 not be subreddit-specific.

3 Methodology

Our job is to solve the text classification challenge. But by building on our earlier work, we broaden the challenge of predicting mental health state to a three-class multi-class classification task.

3.1 Problem Formulation:

We utilize multi-task learning to carry out several classification tasks concurrently, each of which is linked to a distinct set of classes. Each class represents a mental health disorder or a control. (which is a mental health neutral class). The main goal is to build a shared representation that can capture the underlying patterns across several tasks and their related classes, improving performance on each task. The problem can be formulated as follows: Given a set of reddit posts represented by X_i, $y_{i,1}$,...,$y_{i,T}$, where X_i is the feature representation for the i-th example, and $y_{i,t}$ is the label for the t-th task associated with the i-th example, the objective is to learn a function f(x;θ) that can predict the labels for each task given an input x. $y_{i,t} = f(x;\ \theta)$ Our suggested model is made to optimize a multi-objective loss function that simultaneously reduces the losses associated with each task's individual tasks while encouraging the learning of a shared representation across tasks. The regularization term that promotes information sharing between tasks can be used to describe the multi-objective loss function as a weighted sum of the task-specific losses.

For our particular challenge, each task might have some crucial semantic properties, which results in the correlation among these labels. As a result, we think that message transfer between task-specific layers is important.

In this situation, the challenge is to simultaneously strike a balance between task-specific performance and the common representation and to choose the proper weighting for the loss terms. For the message-passing module, a joint probability must also be learned. Additionally, the model architecture must be carefully planned to enable information exchange between jobs while yet enabling task-specific processing. The ultimate objective of multi-class multi-task learning is to enhance the model's performance across a range of activities while minimizing the requirement for task-specific training data and the computational expense of developing separate models for each activity.

We are tackling two significant challenges related to predicting mental health states using Reddit posts. The first challenge is the time-consuming and laborious task of manually labeling data due to the lack of reliable, large datasets. To overcome this, we propose a hybrid active learning system that optimizes the labeling of the unlabeled dataset. The second challenge is understanding the shared latent feature space and task-specific semantic space for semantic information since various mental health issues are interrelated. To address this challenge, we propose a multi-task learning (MTL) framework. Additionally, we introduce a post-hoc explainability module that extracts themes to interpret the key aspects of the task.

3.2 Active Learning Module:

When it comes to minimizing the labeling effort required to curate an initially sizable training dataset, active learning algorithms have proven to be particularly effective.
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. [24] The Least Confidence technique is a common active learning method that chooses instances when the model creates the most entropy, indicating that it is least confident in its predictions. citelewis1994sequential. Since manually identifying data is time- and money-consuming, this method is frequently employed. In this method, a model is first trained on a small dataset before being applied to make predictions on a bigger dataset without labels. For each case in the unlabeled dataset, the result is a collection of probabilities corresponding to each potential class label. Examples for which the model gives the most likely class label a low probability are those for which the author is least certain. These least confident instances are chosen by the active learning technique based on the Least Confidence score and sent to a human annotator for labeling. This enables the model to pick up knowledge from the most instructive cases, which are also the most challenging to categorize. The labeled samples are then

included in the training dataset and used to repeatedly retrain the model until it performs at the desired level or until there is a sufficient amount of data. The Least Confidence technique can assist enhance the model's performance in a straightforward yet effective way by concentrating on the most instructive cases.

- Based on disagreement sampling: The use of several categorization models during training distinguishes this technique from uncertainty-based approaches. We calculate the disagreement for each instance using the vote and consensus probability over all the classifiers following each training iteration. The chosen examples are those where there is the most dispute. To calculate the disagreement, we use the Kullback-Leibler divergence, which can be expressed as

$$D_{\mathrm{KL}}(PQ) = \sum_{i} P(i) \ln \frac{P(i)}{Q(i)} \tag{1}$$

where Q is the consensus prediction of the set of learners and P is the prediction of individual learners. The basic idea behind Disagreement Sampling is that if the models disagree on the prediction of a certain example, then it is likely that this example is difficult to classify and requires additional labeling. By selecting these examples, the model can improve its accuracy by focusing on the most informative examples that are most likely to improve the model's generalization performance.

In this work we initially use submissions from the depression and anxiety subreddit and a model is first trained on a small labeled dataset of 2000 datapoints where an 80-20 split is done. 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.

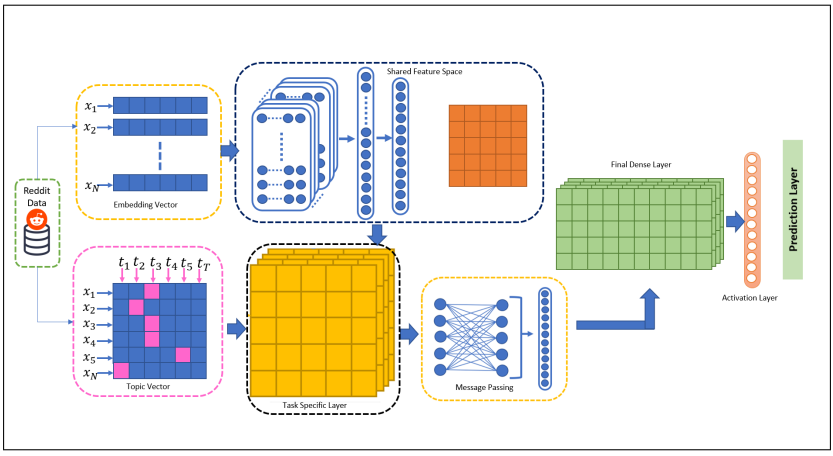

Fig. 2 The illustrative architecture of the proposed GMMTL method.

3.3 N-gram Feature Extraction:

N-grams are sequences of N words or characters, and they can be used as a feature extraction method to capture a mixture of 1, 2, or 3-word sequences as tokens. For instance, a unigram model (n=1) would represent the sentence "I am not interested in living anymore" as tokens of 1 word, such as ["interested", "in", "living", "anymore"] after the removal of stopwords. On the other hand, a bigram model (n=2) would represent the same sentence as tokens of 2 words, such as ["not interested", "interested in", "in living", "living anymore"], which captures the context of the text. By using bigram features, the model can recognize the phrase "not interested" as a single entity, which would not be possible with a unigram model. Therefore, we hypothesize that N-grams should be used as a precursor to our feature extraction pipeline.

Topic Modeling is an unsupervised machine learning technique that represents the topic categories with which a set of text can be represented. It can divide the corpus into a fixed number of topics, each containing the most important words or phrases. Additionally, it provides a token-level representation of a topic where each token is given a probability score for belonging to different topics. In our work, we consider each submission a single text document, and we are interested in the probabilistic distribution of different topic categories for each text sample 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[25, 26].

3.4 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 3.3 and b) topic distribution vector based on the topic modeling technique mentioned in subsection 3.3. Since a sentence-level embedding serves as a general feature representation for a text corpus, we feed this set of features into a shared component consisting of interconnected hidden layers. To separately learn the topic-level representation specific to each post, we employ a task-specific feature learning module.

Shared Feature Learning Module: The model in Fig2 is based on CNN architecture. It starts with an embedding layer that represents each post's word-level embedding using 200 dimensions. This embedding layer utilizes a pretrained BERT-base model, which is more effective at capturing contextual lexical features for each word in a document. Next, a convolutional layer is employed with 64 filters and a filter size of 5, followed by dropout. The layer's output undergoes max-pooling and a fully connected dense layer, acting as a feature-sharing space for all tasks. The module's output is a tensor of length 32, which is concatenated with the Task-specific topic representation vector

which is then fed into Task-specific feature learner. This design is particularly suited to capture the long-form nature of Reddit submissions.

Task-specific Feature Learner: For the problem of muti-task text classification, an important objective is to learn the task-specific features. To illustrate, when discussing anxiety in posts, specific terms such as "Xanax," "stress," "fidgety," and "racing mind" are typically used, whereas when identifying signs of depression, terms like "die," "want to die," and "lonely" are more prevalent. To establish this correlation between task and topic, we utilize unsupervised topic modeling techniques. Hence, as input of this module for each sample we concatenate the output of the shared module with 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 concatenated vector $\mathbf{T} \in \mathbb{R}^{n \times d}$ where n is the number of instances in the input and d denotes the dimension after concatenation. For a fully connected dense layer of m number of nodes, the vector is multiplied by the weight vector of dimension $(d \times m)$ and the output of the layer can be written as follows:

$$\mathbf{H} = \mathbf{XW} + bias \tag{2}$$

The output of this layer is thus a tensor of size $n \times p$. For each task we train this layer separately allowing it to learn task-specific features individually.

Message Passing Layer: We propose a message passing layer to spread the learned features between different tasks. This helps the model learn from multiple related tasks simultaneously. The graph structure of the layer is described as follows:

Let G = (V, E) be an undirected graph with vertices V and edges E. V = v_1, v_2, ..., v_n where v_i is task i. E= e_{12}, e_{23}, ..., e_{mn} where e_{ij} represents an interaction between node i and j. X = x_1, x_2, ..., x_n are input features for all task-specific modules. The hidden representation learned after the first pass through the task-specific module is H_i. The following equation is the aggregation method:

$$\mathbf{H}_i = \mathbf{U}_i \mathbf{x}_i + \sum_{j \epsilon N(i)} \mathbf{W}_{ij} \mathbf{h}_j \tag{3}$$

where U_i is the weight matrix for the task i's input features and N(i) is the set of neighboring tasks of task i in the graph, and W_{ij} is the weight matrix for the edge between task I and task j. The hidden representations h_i of all tasks are then combined to form a joint probability distribution p_{joint} overall tasks. Here we use a fully connected graph with each node representing a task.

3.5 Computing Overall Loss Function for joint learning:

The overall loss function in each epoch includes computing both the binary cross-entropy loss and the joint probability derived from the message passing module. 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:

$$\mathbf{E}_{loss} = \sum_{j=1}^{\mathbf{n_{task}}} \alpha_j(\mathbf{E}_{BCELoss}(\hat{Y}, Y) + \beta \mathbf{L}_j \tag{4}$$

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 functions, and β is another hyperparameter designed to control the weightage of the joint loss function L_j which is the negative log likelihood of the joint probability distribution p_{joint} derived from the message passing module. The joint probability term encourages the model to learn relationships between tasks and improve performance on related tasks. Our objective is to minimize the loss while also updating the model's parameters until they converge.

3.6 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 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 joint probability term that is applied to the cross entropy loss is derived from the message passing mechanism that is created based on the graph structure. Next, the hidden representation from the shared layer is concatenated with the extracted topic feature vector from the topic model and passed through the task specific dense layers. At this point, we use the initial input to the task-specific module and output of each task specific module to do the computaion for message passing which further generates the joint probability loss. Finally we compute each task's cross entropy loss function and add that to the joint loss function to get the overall loss function. We minimize the loss function by performing backpropagation on the weights of each layer according to its gradient and learning rate.

.

4 Experiments

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

4.1 Datasets:

Reddit discussions on a particular topic take place in subreddits denoted by the prefix 'r/'. Our focus is on the submissions within these subreddits, which consist of a title and body of variable length. Our data collection spans 'r/MentalHealth', 'r/SuicideWatch', 'r/Anxiety', 'r/bipolar', and 'r/BPD'. Previous approaches assigned labels based on the subreddit topic (e.g., depression-specific labels for depression subreddits and anxiety-specific labels for anxiety

Table 1 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)

Emb	**Logistic**			**KNN**			**SVM**			**Random Forest**			**MLP**			**GMMTL**		
2-22	**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	0.708	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	0.713	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	0.762	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	0.761	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	0.752	0.763	0.757	0.851	0.839	0.845	0.758	0.773	0.765	0.875	0.861	0.868	0.872	**0.889**	0.880
BERTopic+BERT	0.785	0.771	0.778	0.745	0.728	0.736	0.879	0.863	0.869	0.779	0.765	0.772	0.873	0.859	0.866	**0.889**	0.882	**0.885**
EMPATH+BERT	0.744	0.757	0.750	0.732	0.725	0.728	0.811	0.823	0.817	0.759	0.741	0.749	0.825	0.811	0.866	0.851	0.837	0.843

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 $D_{\mathrm{KL}}(PQ) = \sum_i P(i) \ln \frac{P(i)}{Q(i)}$ 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:$W, \Theta, G(V, E)$;

while t <= epoch **do**
  $H_{shared} = F_{\Theta_1}(X_{embedding})$
  while i <= tasks **do** $x = H_{shared}$*X_t where X_t is the topic representation vector
    $H = X * W + bias$
    $H_i = \mathbf{Dense}(H)$ $H_i = U_i * x_i + \sum_{j \epsilon N(i)} W_{ij} \mathbf{h}_j$ Message passing
    $\hat{Y}_i = \mathbf{Dense}_{\Theta}(H_i)$
  end while
  $E_{loss} = \sum_{j=1}^{n_{task}} \alpha_j (E_{BCELoss}(\hat{Y}, Y) + \beta L_j)$
  Loss.backward()
  Update parameters
end while

Table 2 GMMTL fine-grained results

Category	**Percision**	**Recall**	**F1-Score**
Depression	0.913	0.874	0.893
Anxiety	0.872	0.899	0.885
Other	0.860	0.873	0.866

subreddits). However, it is common to find indications of different mental health conditions in these subreddits. For instance, 'r/Anxiety' may contain posts related to both anxiety and depression, as do other subreddits. Therefore, we aim to curate a dataset of 2000 submissions with two labels, 'Anxiety'

Table 3 Active Learning training of GMMTL 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 4 Most Important Topics for Depression Detection

Topic id	Top Phrases/Words
4	"depression" " feel" "depressed" "feeling" "want" "get" "life" "really" "even"
23	"thoughts" "mind" "things" "intrusive"
9	"want" "to die" "dead" "life" "want dead" "life"
24	"help" "suicidal" "suicidal hotline" "hotline" "need" "want""talk "feel"
10	"therapy" "therapist" "appointment" "need" "help" "cbt" "know" "get"

Table 5 Most Important Topics for Anxiety Detection

Topic id	Top Phrases/Words
8	"feel" "know" "anxiety" "depression" "want" "get" "life" "really" "time"
3	"side effect" "zoloft" "lexapro" änyone" "meds" "take"
23	"thoughts" "mind" "things" "intrusive"
14	"work" "job" "home" "go" "day" "covid
6	änxious" "feeling" "calm" öften" "lot" "älso" älways" "worrying"

and 'Depression'. Then, to train the model on non-mentalhealth related data we collect data from subreddits like 'r/CasualConversation','r/cozyplaces' etc. The data from these subreddit are labeled as control and are included in the training dataset to train the model to identify not notly anxiety and depression related post but also to differentiate them from the unrelated posts found in control group.

Our data collection from Reddit focused on the most frequently visited subreddits related to depression and anxiety during the period spanning from January 2020 to January 2022. To extract all the submissions from these subreddits, we utilized PushShift.io [5]. After obtaining the dataset, we merged

the data from various subreddits into a single dataset in preparation for fine-grained labeling of an initial set of 5000 data points.

4.2 Experiment settings & Baselines

Both topic modeling and word embeddings are essential techniques for text classification [25, 27–29]. Thus, in our experiments, we employ a combination of features derived from both word embeddings and topic modeling across all base models. We also use linguistic dictionaries such as EMPATH to convert text data into feature vectors. For word embeddings, we utilize the pre-trained BERT model [30], and for topic modeling, we primarily use **LDA** and **BERTopic**. **LDA** is a well-known unsupervised topic modeling technique that has been widely used in mental health research and has demonstrated superior performance compared to other dictionary-based methods. On the other hand, **BERTopic** is an unsupervised topic modeling technique that employs BERT-based word embeddings to create clusters based on UMAP and HDBSCAN and has been gaining popularity in recent times.

In addition, we focus on statistical methods that 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 corpus, we use **TF-IDF** features. The feature value for each token is calculated by computing the term and inverse document frequency [31].

We use 80 percent of the labeled data as the training dataset and employ a K-Fold cross-validation setting with a training and validation split of 80:20 during training for all the base models. Table 1 shows the average performance across all metrics after 10 independent runs. For our proposed Graph-Based Multitask Learning(GMMTL) MentalHealth condition prediction model, we train the model incrementally from 2000 to 5000, with a pool of 300 data points being manually labeled at each iteration. Table 3 shows the model's performance at each iteration on the 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 efficient in any classification task. It models the probability of a discrete outcome given an input variable.
- **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.
- **KNN CLlassifier:** This is a KNN based classifer.

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

4.3 Results and Discussion:

Table 1 details results across different experimental settings using F-1, Recall, and Precision metrics. The models were all trained on the curated labeled dataset of 5000 data as explained in subsection 4.2. To create this 5000 labeled dataset we use active learning with classification for only depression and anxiety (which is why the reported accuracy in Table 3 are not indicative of the performance of the final model. Our hypothesis was that if we start with a small set of human annotated data and in each iteration only label data that are informationally rich, then we may not need large dataset of more than tens of thousands for training. It has the potential to save cost in both labeling and data training where smaller dataset also leads to a less computaionally heavy training process. To this effect, we employ the disagreement based active learning method where a set of different models were trained on the initial dataset of 2000. In this stage, since at each iteration we are training multiple models, to alleviate the training cost we use lightweight machine learning models in the form of SVM, Random Forest, MLP and logistic regression. In Table 3 we detail the average accuracy scores of the these models on each iteration. From the initial observation, it is clear that the accuracy of the model improves steadily, more or less. To decide when to stop this process we rely on the disagreement score based on Kullback-Leibler divergence value. When this value for the least confident datapoints from the unseen pooled dataset starts plateauing we stop the process. In our case we stop when the dataset length is 5000. We then added the control dataset where each datapoint from those subreddits are labeled negatively for both deression and anxiety and are given positive value for the label "control".

Then we apply GMMTL our current modified version with the aforementioned baseline models on the three class multilabel classification task. 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 on LDA+BERT and BERTopic+BERT experimental settings across all models. To compare the performance of topic model extracted features to predefined dictionary based methods we use EMPATH as an additional feature [32]. EMPATH is a lexcial category genertaed by pretrained deep neural network model that has been widely used to interpret documents and its lexical categories. We use the lexical categories generated by EMPATh to use them in the same vain as the topic distribution vector where each document is expressed as fixed length vector of dimensions the size of categories. Our experiments prove that despite being more effective than TF-IDF, it still is not abl to outperform the feature vector of BERT+Topic(LDA/BERTopic). This could be down to the sparse nature of the vector as long submissions of Reddit will result in lot of categories having zero values. On the other hand, results with 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 multitask-learning model with message passing mechanism. It can be inferred from the table that GMMTL produces the best scores in most of the metrics. However, SVM and MLP models do come close in terms of Precision.

In Table 2 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. However, we also see how the performance on the control group is not that satisfactory especially when predicting out of class. This could be due to the significant overlap in certain topics between a more "neutral" text and a more "domain specific" text. .

5 Case Study

5.1 Motivation:

The main motivation behind conducting case study into the explainable features corresponding to each task is the initial data analysis. Interestingly we observed some key similarity and differences in wordcould generated for both sets of data for depression and anxiety. This led us to question how can we interpret the predictive model by giving explanation on the topic level? As topics are consisted of a set of words, it provides a perfect opportunity to understand what are the most contributing features in both the classes. We wanted observe the intersectionality and points of differneces between the two categories. In our previous study we employed a group lasso regularization term to get anti-hoc explainability. However, because of the design of the current architecture where the concatenation takes place before feeding it into

task-specific module and the introduction of message passing induced joint probability distribution based loss, the group lasso regularing term will not have the same effect. Hence, we rely on post-hoc explainability.

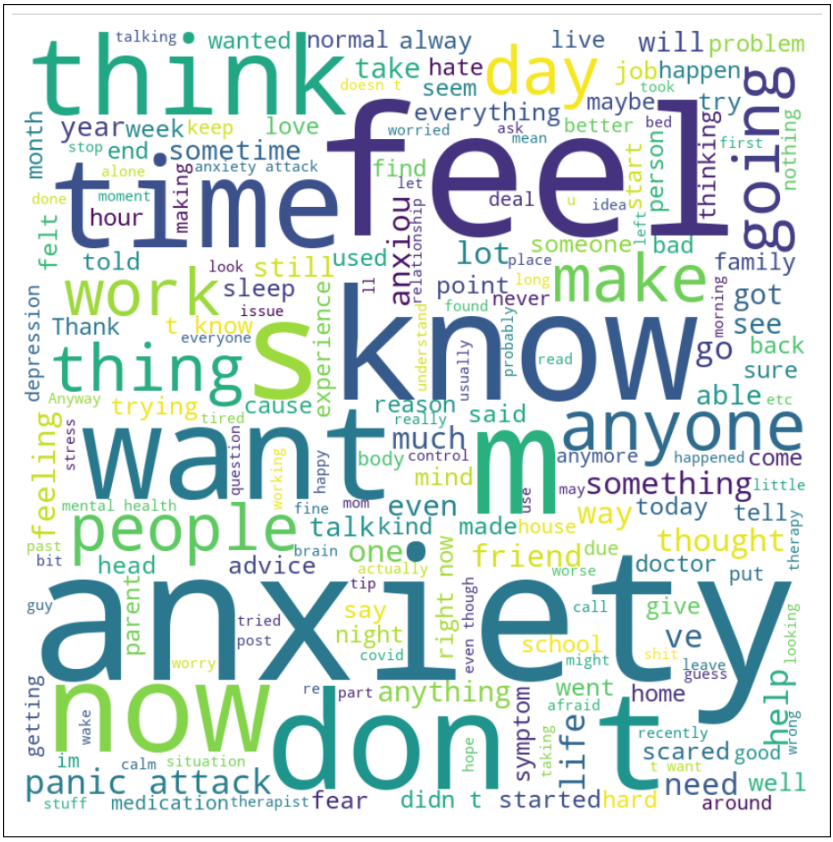

Fig. 3 Anxiety Wordcloud

5.2 Most important Topics Discovery:

The corresponding Topic Modeling technique gives significant insight into these topics by listing the top keywords for each topic. As shown in Table5 and Table 4, there are some intersectionality in the most important topic s for prediction of both the classes. The Topic id 23 consists of words like "thoughts", "mind", "things", "intrusive". The result also shows that documents having these topic and words are often classified as both depression and anxiety. This is potentially due to the co-occurrence of both these conditions in a lot of people that often gets expressed through these long form posts. Another important observation for anxiety prediction is that topics that include speciic medicine

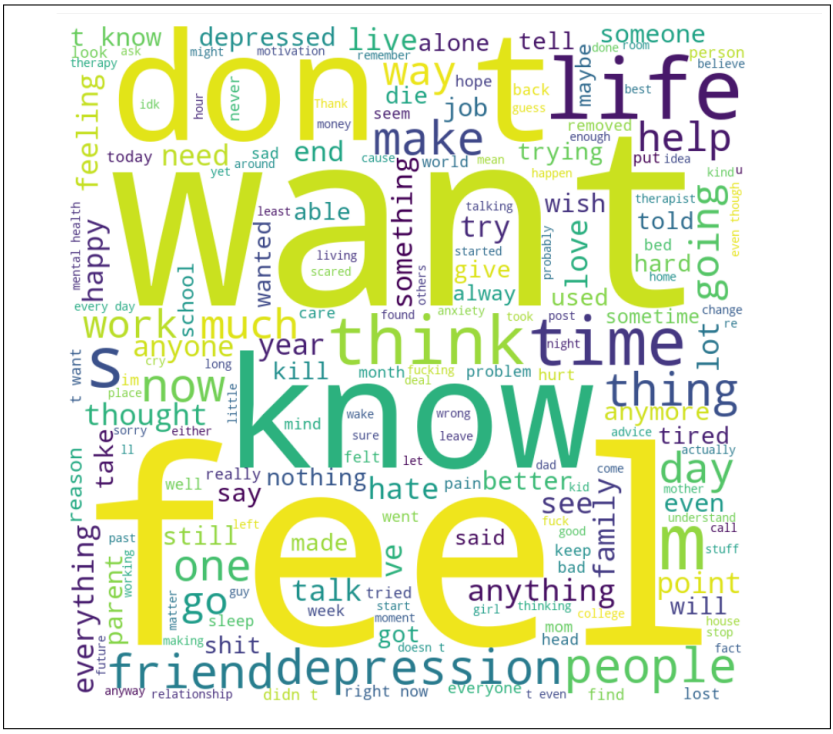

Fig. 4 Depression WordCloud

names can often be indicator of the existence of conditions like anxiety disorder. Topic id 3 for example contains both zoloft, lexapro which are often assosciated with people suffering from some kind of anxiety disorder. Another interesting observation is the existence of the topic related to COVID-19. As the data contains posts from 2020 onwards, it is a valuable insight to notice how in anxiety related conversation COVID-19 related discussions are prevalent. On the other hand, for depression topic that is more telling for detection s has words like "depression", "die","suicide"(Topic 9). However, one interesting observation is that lot of the topics that have words like "die" or "suicide", also includes words like "hotline", "help" which indicates a severe mental health crisis. This could be particularly useful for mental-health crisis workers who can use the detection tool and additionally identify severe cases such as this to identify or offer help to people in need.

6 Conclusion

In this paper, we proposed a novel Message-passing Multi-task learning model GMMTL, that extracts task-specific features in the form of topics and learns from a shared feature space of document-level embedding. Additionally, the graph-based message passing mechanism enables the model to learn correlation among different tasks. Our framework expands our previous work on mental

health prediction from a mentalhealth related subreddit-specific approach to a more general data space by adding control population through the means of other subreddits that are not related Reddit. We successfully show that the combination of word embeddings and topic modeling can be leveraged even further through the message passing scheme. We tackle the problem of labeling cost through a more detailed description of the adopted active learning approach which shows even with comparatively less amount of data a good model can be trained for such task. 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 that improves on the previous contribution by giving class-specific explanation. The insights collected from this has the potential to be extremely valuable to social worker or mental-health crisis volunteers. We substantiate our findings through extensive experiments. A future direction of this work could be to look into specific mental disorders like "ADHD" "BPD," or "bipolar" and try to predict or detect such conditions with explainability, as shown in this paper.

References

[1] Sarkar, S., Alhamadani, A., Alkulaib, L., Lu, C.-T.: 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), pp. 427–435 (2022). IEEE

[2] Dor, L.E., Halfon, A., Gera, A., Shnarch, E., Dankin, L., Choshen, L., Danilevsky, M., Aharonov, R., Katz, Y., Slonim, N.: Active learning for bert: an empirical study. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 7949–7962 (2020)

[3] Seung, H.S., Opper, M., Sompolinsky, H.: Query by committee. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory, pp. 287–294 (1992)

[4] Holub, A., Perona, P., Burl, M.C.: Entropy-based active learning for object recognition. In: 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 1–8 (2008). IEEE

[5] Yang, Y., Ma, Z., Nie, F., Chang, X., Hauptmann, A.G.: Multi-class active learning by uncertainty sampling with diversity maximization. International Journal of Computer Vision **113**(2), 113–127 (2015)

[6] Ji, T., Fu, K., Self, N., Lu, C.-T., Ramakrishnan, N.: Multi-task learning

for transit service disruption detection. In: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 634–641 (2018). IEEE

[7] Zhang, Y., Yang, Q.: A survey on multi-task learning. IEEE Transactions on Knowledge and Data Engineering (2021)

[8] Liu, P., Fu, J., Dong, Y., Qiu, X., Cheung, J.C.K.: Learning multi-task communication with message passing for sequence learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 4360–4367 (2019)

[9] Zhao, L., Chen, J., Chen, F., Jin, F., Wang, W., Lu, C.-T., Ramakrishnan, N.: Online flu epidemiological deep modeling on disease contact network. GeoInformatica **24** (2020). https://doi.org/10.1007/s10707-019-00376-9

[10] Comito, C.: Sensing social media to forecast covid-19 cases. In: 2022 IEEE Symposium on Computers and Communications (ISCC), pp. 1–6 (2022). IEEE

[11] Jahanbin, K., Jokar, M., Rahmanian, V., *et al.*: Using twitter and web news mining to predict the monkeypox outbreak. Asian Pacific Journal of Tropical Medicine **15**(5), 236 (2022)

[12] De Choudhury, M., De, S.: Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In: Eighth International AAAI Conference on Weblogs and Social Media (2014)

[13] Tsugawa, S., Kikuchi, Y., Kishino, F., Nakajima, K., Itoh, Y., Ohsaki, H.: Recognizing depression from twitter activity. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pp. 3187–3196 (2015)

[14] Coppersmith, G., Leary, R., Whyne, E., Wood, T.: Quantifying suicidal ideation via language usage on social media. In: Joint Statistics Meetings Proceedings, Statistical Computing Section, JSM, vol. 110 (2015)

[15] Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K.: 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, pp. 1–10 (2015)

[16] Preotiuc-Pietro, D., Carpenter, J., Giorgi, S., Ungar, L.: Studying the dark triad of personality through twitter behavior. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, pp. 761–770 (2016)

[17] De Choudhury, M., Kiciman, E., Dredze, M., Coppersmith, G., Kumar, M.: 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, pp. 2098–2110 (2016)

[18] Eichstaedt, J.C., Schwartz, H.A., Kern, M.L., Park, G., Labarthe, D.R., Merchant, R.M., Jha, S., Agrawal, M., Dziurzynski, L.A., Sap, M., *et al.*: Psychological language on twitter predicts county-level heart disease mortality. Psychological science **26**(2), 159–169 (2015)

[19] Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Lucas, R.E., Agrawal, M., Park, G.J., Lakshmikanth, S.K., Jha, S., Seligman, M.E., *et al.*: Characterizing geographic variation in well-being using tweets. In: Seventh International AAAI Conference on Weblogs and Social Media (2013)

[20] Jaidka, K., Giorgi, S., Schwartz, H.A., Kern, M.L., Ungar, L.H., Eichstaedt, J.C.: 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)

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

[22] Macalli, M., Navarro, M., Orri, M., Tournier, M., Thiébaut, R., Côté, S.M., Tzourio, C.: A machine learning approach for predicting suicidal thoughts and behaviours among college students. Scientific reports **11**(1), 1–8 (2021)

[23] Sawhney, R., Joshi, H., Gandhi, S., Shah, R.R.: Towards ordinal suicide ideation detection on social media. In: Proceedings of the 14th ACM International Conference on Web Search and Data Mining, pp. 22–30 (2021)

[24] Lewis, D.D., Gale, W.A.: A sequential algorithm for training text classifiers. In: SIGIR'94, pp. 3–12 (1994). Springer

[25] Wei, X., Croft, W.B.: 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, pp. 178–185 (2006)

[26] Grootendorst, M.: Bertopic: Neural topic modeling with a class-based tf-idf procedure. arXiv preprint arXiv:2203.05794 (2022)

[27] Mo, Y., Kontonatsios, G., Ananiadou, S.: Supporting systematic reviews using lda-based document representations. Systematic reviews **4**(1), 1–12 (2015)

[28] Wahid, J.A., Shi, L., Gao, Y., Yang, B., Tao, Y., Wei, L., Hussain, S.: Topic2features: a novel framework to classify noisy and sparse textual data using lda topic distributions. PeerJ Computer Science **7**, 677 (2021)

[29] Schick, T., Schütze, H.: Bertram: Improved word embeddings have big impact on contextualized model performance (2020)

[30] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding, 4171–4186 (2019). https://doi.org/10.18653/v1/N19-1423

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

[32] Fast, E., Chen, B., Bernstein, M.S.: Empath: Understanding topic signals in large-scale text. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pp. 4647–4657 (2016)