Journal of Biomedical Informatics 150 (2024) 104586

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Original Research

MULTIWD: Multi-label wellness dimensions in social media posts

Muskan Garg [a,*], Xingyi Liu [a], M.S.V.P.J. Sathvik [b], Shaina Raza [c], Sunghwan Sohn [a]

[a] Mayo Clinic, Rochester, 55901 MN, USA
[b] IIIT Dharwad, Goa, 580011 IN, India
[c] Vector Institute for Artificial Intelligence, Toronto, M5G 1M1 ON, Canada

ARTICLE INFO

Keywords:
Dataset
Mental health
Multi-label classification
Wellness dimensions

ABSTRACT

Background: Halbert L. Dunn's concept of wellness is a multi-dimensional aspect encompassing social and mental well-being. Neglecting these dimensions over time can have a negative impact on an individual's mental health. The manual efforts employed in in-person therapy sessions reveal that underlying factors of mental disturbance if triggered, may lead to severe mental health disorders.
Objective: In our research, we introduce a fine-grained approach focused on identifying indicators of wellness dimensions and mark their presence in self-narrated human-writings on Reddit social media platform.
Design and Method: We present the MULTIWD dataset, a curated collection comprising **3281** instances, as a specifically designed and annotated dataset that facilitates the identification of multiple wellness dimensions in Reddit posts. In our study, we introduce the task of identifying wellness dimensions and utilize state-of-the-art classifiers to solve this multi-label classification task.
Results: Our findings highlights the best and comparative performance of fine-tuned large language models with fine-tuned BERT model. As such, we set BERT as a baseline model to tag wellness dimensions in a user-penned text with F1 score of 76.69.
Conclusion: Our findings underscore the need of trustworthy and domain-specific knowledge infusion to develop more comprehensive and contextually-aware AI models for tagging and extracting wellness dimensions.

1. Background

The United Nations' (UN) "Transforming our World: the Agenda 2030 for Sustainable Development" resolution adopted in September 2015 [1] presents a comprehensive vision for addressing the Sustainable Development Goals (SDGs). The third SDG, "Ensure healthy lives and promote well-being for all at all ages" aims to reduce premature mortality from non-communicable diseases by one-third by the year 2030 [2]. This ambitious goal highlights the importance of prioritizing and improving global health and overall well-being. Moreover, according to the World Health Organization (WHO), approximately one in four people worldwide will experience a mental health issue at some point in their lives.[1] This mental disturbance is caused by genetics, environment, and other life experiences. Mental disorders, have significant impact on an individual's overall wellbeing, including their ability to work, socialize, and carry out daily activities, resulting in physical health problems, such as an increased risk of cardiovascular disease, diabetes, and obesity.

Dunn's model of wellness dimensions is a conceptual framework that describes wellness as a multidimensional and holistic concept [3]. The model consists of six dimensions of wellness or aspects,[2] including physical, emotional, social, intellectual, spiritual, and occupational dimensions. As the concept of wellness is a holistic approach to health that encompasses various dimensions of well-being, maintaining balance in all these dimensions is crucial for achieving mindfulness and cognitive balance [4]. Any neglect or disregard for any *wellness dimension* can have adverse effects on an individual's mental health and cognitive function, leading to cognitive decline [5]. For instance, neglecting the 'social' dimension of wellness can lead to loneliness and isolation.

1.1. Objective

As a starting point of our study, we approach the problem of identifying multiple wellness dimensions discussed in a Reddit post

* Corresponding author.
E-mail addresses: garg.muskan@mayo.edu (M. Garg), liu.xingyi@mayo.edu (X. Liu), 20bec024@iiitdwd.ac.in (M. Sathvik), shaina.raza@vectorinstitute.ai (S. Raza), sohn.sunghwan@mayo.edu (S. Sohn).

[1] https://www.who.int/health-topics/mental-health#tab=tab_1.

[2] Note that *aspect* is another term given to the *dimension*.

https://doi.org/10.1016/j.jbi.2024.104586
Received 12 October 2023; Received in revised form 6 December 2023; Accepted 3 January 2024
Available online 6 January 2024

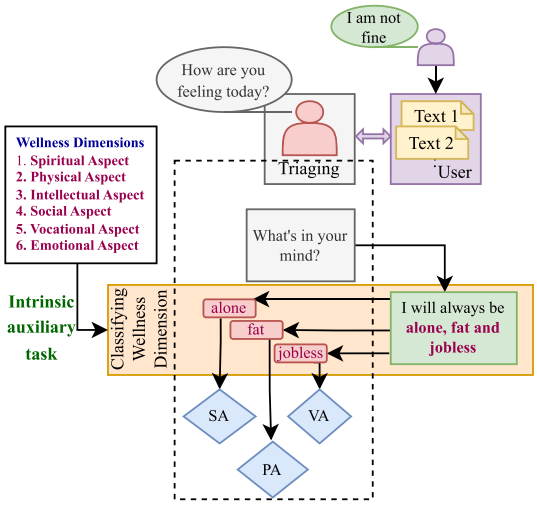


Fig. 1. Overview of the Mental health triaging and pre-screening application. Wellness dimensions as an intrinsic auxiliary task. Figure represents three different aspects present in one single sentence. Here SA: Social Aspect, PA: Physical Aspect, VA: Vocational Aspect.

as a multi-label classification task. To accomplish this, we construct and release an English Reddit social media dataset that allows for the development of comprehensive and contextual AI models to determine all the indicators of impacted wellness dimensions in a given post. Consider the example post *P*:

> *P*: I'm 21 years old. I have aspergers syndrome and depression← (Physical Aspect), I have struggled quite a lot and I want to do stuff my own way to get better (with the help from actual professionals). My mum, dad and step-mum← (Social Aspect) won't leave me alone and they constantly make choices for me and it's starting to get to me. They make me feel unhappy and miserable← (Emotional Aspect). What should I do?

A 21 years old user mentions about their *Asperger's syndrome and depression*, which is indicative of the medical problem, closely associated with the impact on physical aspect of wellness. The user mentions that their interpersonal relations such as *mum, dad, and step-mum* do not give them personal space, impacting user's sentiments and social aspect of wellness. Finally, the user mentions feeling *unhappy and miserable* due to their parents' interference, which affects their emotional aspect. Thus, we can develop AI models that are better equipped to identify and emphasize particular aspects of deteriorating mental health by marking text-spans specifying discussion about multiple wellness dimensions in a given post (see Fig. 1).

1.2. Our contributions

In this paper, we present following major contributions that advance the understanding of multiple wellness dimensions in social media text:

1. We define a fine-grained task of identifying wellness dimensions in human writing to aid the concept extraction for early detection of chronic mental disturbance against wellness during mental health triaging.
2. We construct and release MULTIWD, a corpus of 3281 instances, annotated to identify multiple wellness dimensions in a given Reddit post.
3. We evaluate the performance of traditional multi-label classifiers on MULTIWD to establish baselines and find that there is significant room for improvement.

To the best of our knowledge, this is the first of its kind, quantitative study to introduce the need of determining dimensions of wellness to bolster existing studies on mental disturbance in Reddit posts.

The clinical psychologists who provide in-person therapy sessions notice that social isolation, poor hospitalization, and poor daily activities decline wellness. However, with the growing demand for mental healthcare and limited availability of mental health professionals, timely and quality care becomes a major challenge. As such, there is a need to simulate the process of identifying the impacted wellness dimensions at an early stage. It is essential to identify which dimensions of wellness are impacted in an individual's expression of mental disturbance, to aid the development of personalized and targeted interventions, for example, an individual with emotional disturbance might require interventions focused on coping mechanisms and stress management.

Anonymity on social media makes it easier to share their mental health concerns on social media when they are hesitant with in-person therapy session due to privacy or social stigma concerns. The social media allows for a wider reach than in-person therapy sessions, making it easier and beneficial for individuals who lack access to in-person mental health services.

Statement of Significance

Summary	Description
Problem	Identifying aspects of wellness, rather than illness, in texts written by users on social media to understand their perceptions and psychological state. This approach is taken before exploring broader theories like Social Determinants of Health (SDoH).
What is already known	A well-established Dunn's High Level Wellness theory.
What this paper adds	This study extract the information by tagging and labeling wellness dimensions (through existing Dunn's High level wellness theory), that is discussed in a user-penned text to aid the extraction of users' perception through AI models.
Applications	We observe a progression from confusion to political discourse, and finally to existential thoughts, highlighting the impact of experiences on various wellness dimensions and supporting cohort studies. Our dataset, designed to identify these changes, can assist in early intervention of crucial factors such as identifying at-risk individuals who may need support, thereby contributing towards mental health pre-screening and triaging.

2. Related work

While progress has been made in developing classifiers, there is a lack of research and developments in fine-grained analysis. For instance, Coppersmith et al. [6] conducted a study examining the language patterns of individuals on social media platforms and found that the usage of specific words such as "alone", "empty", and "depressed" indicate a higher likelihood of suicide risk. Subsequently, the research community began exploring the use of machine learning algorithms to analyze social media data for mental health analysis to identify individuals with depression on Twitter by analyzing their linguistic patterns [7–9]. However, mental health is multi-faceted and includes numerous wellness dimensions. To offer a broader perspective on mental well-being, our study introduces a detailed approach that incorporates wellness against illness.

The historical evolution of language resources on mental disturbance in social media posts has seen significant advancements in recent years. In addition to the traditional tasks, the research community has increasing concerns with fine-grained tasks such as identifying the underlying causes of mental health issues [10,11], fostering the need of explicitly considering wellness dimensions as an intrinsic fine-grained task.

Dunn's model of wellness dimensions focuses on individual behavior and personal wellness, making it suitable for analyzing texts related to personal health practices and wellness attitudes. In contrast, existing models such as SDoH are more effective for analyzing texts concerning community health, social policies, and environmental factors, as shown in existing studies [12,13]. Dunn's model emphasizes a holistic view of health by quantifying personal writings, suggesting a focus on personalized wellness rather than illness. This approach contributes to personal health management, self-care practices, and individual health choices. However, the SDoH 2030 model is useful for identifying discussions related to policy and structural influences on many external factors, such as healthcare policies, social inequalities, and environmental issues, which cannot be inferred from cross-sectional users' postings on social media. To bridge this gap, we first determine the wellness area in which a user might need help, followed by fine-grained analysis of infusing external factors through electronic health records (EHR) or other sources of information. Moreover, the type of user-generated text being analyzed could influence which model is more applicable, for instance, personal blogs/diaries/social media posts align more with Dunn's model, while forum discussions on social issues can be better analyzed through the SDoH lens.

2.1. High-level wellness theory

Unlike many models that primarily focus on the absence of disease or illness, Dunn's model emphasizes positive health and well-being. Dunn's model places a significant emphasis on personal responsibility and the active pursuit of better health, encouraging research into self-care practices, personal health management, and lifestyle choices that contribute to overall wellness. Dunn's model encourages research into preventive health measures particularly useful in epidemiological studies, health behavior research, and in developing interventions that address more than just physical symptoms. The holistic nature of Dunn's model invites interdisciplinary research, integrating insights from psychology, sociology, public health, and medicine. This can lead to more comprehensive and effective health solutions.

The high-level wellness theory, proposed by Halbert L. Dunn in 1961, suggests wellness as a multidimensional concept that goes beyond mere absence of disease or illness [3,14]. Achieving high-level wellness requires a *proactive and preventive* approach to health care, as opposed to a *reactive* approach focused on treating illness or disease. According to Dunn, wellness is a state of optimal health and well-being that is achieved through the balance of following six different dimensions physical, emotional, social, intellectual, vocational, and spiritual dimensions [15,16]. These dimensions are interconnected and together contribute to an individual's overall well-being.

Spiritual Aspect (SpA): A wellness dimension that emphasizes discovering meaning and purpose in life through practices like meditation, prayer, yoga, and connecting with nature or cultural rituals, fostering inner peace and mindfulness.

Physical Aspect (PA): Physical development involves body growth and adopting healthy habits, such as a balanced diet and avoiding harmful substances, while body shaming can increase self-awareness about appearance and health, potentially leading to negative emotions like shame, self-doubt, and depression.

Intellectual Aspect (IA): A wellness dimension that focus on fostering intellectual growth and cultural engagement through activities like reading, lectures, discussions, and creative hobbies, both in academic settings and beyond, to enhance cognitive understanding of the world.

Social Aspect (SA): A wellness dimension that highlights the interplay between individual, societal, and environmental well-being, emphasizing the importance of social connections, cultural appreciation, and strong family and community ties for fostering a sense of belonging and support.

Vocational Aspect (VA): Occupational fulfillment contributes to life satisfaction by fostering creativity, professional growth, and financial management, but it also acknowledges the risks of work-related stress, burnout, and financial strain that can adversely affect mental well-being.

Emotional Aspect (EA): The wellness dimension of emotion centers on comprehending and embracing one's own emotions. Based on Plutchik's wheel framework, 13 emotion labels are used, grouped into opposing pairs: Love vs. Hate, Joy vs. Sadness, Trust vs. Disgust, and Anticipation vs. Surprise, as referenced in Kumar's 2022 study [17]. Emotional wellness, according to Dunn's model, encompasses a wide range of emotional states and their impacts on overall life satisfaction and personal fulfillment. It is about achieving a balance in life and feeling content and fulfilled as compared to the sentiment analysis that categorize text (like social media posts, reviews, etc.) into positive, negative, or neutral sentiments, without delving into the complexities of emotional states or their impact on an individual's life.

By employing the definitions provided in this section, we can discern the presence of a specific aspect in a person's writings when they describe it. For instance, in order to comment about the presence of given dimension D where $D = \{\text{SpA}, PA, IA, SA, VA, EA\}$, consider the following markings:

Marking for D:

0: Not mentioned

1: The text contains indicators describing D.

As we examine the dimensions of wellness from user-penned social media posts against illness, we track the generalized aspects as compared to the clinical concepts. We observe the potential oversimplification of health issues when using Dunn's wellness model, especially if all medical symptoms and health conditions are aggregated solely under the "physical" aspect. However, understanding the full spectrum of health, especially in a world where mental and social health are increasingly recognized as critical components of overall well-being, is important. Thus, along with phrases signifying the physical aspect, we track other aspects including spiritual and emotional, specifying mental well-being. Building upon the foundational work of explainable multi-class wellness dimension problem, we presents a two-stage dataset specifically designed for a teacher-student model [15,18], enhancing the process of knowledge distillation for wellness dimensions. Our work of multi-label classification is adaptable to real-time pre-screening of participant, contributing towards motivational interviewing for wellness against illness.

3. Design and method

Problem definition: Our research aims to find and label the presence (1) or absence (0) of each of the six pre-defined wellness dimensions in user-penned text, a task made challenging by its subjective, domain-specific and complex nature. As such, we design annotation scheme and develop a multi-label classification method to find a vector of wellness dimensions.

3.1. Corpus collection

Simple labeling can lead to inaccuracies. Thus, we emphasize a consistent annotation process for dataset reliability. We formed an expert panel of a clinical psychologist, rehabilitation counsellor, and social NLP researcher. Their diverse expertise contributed to collaborative annotation guidelines, ensuring accurate and consistent annotations after intensive team discussions.

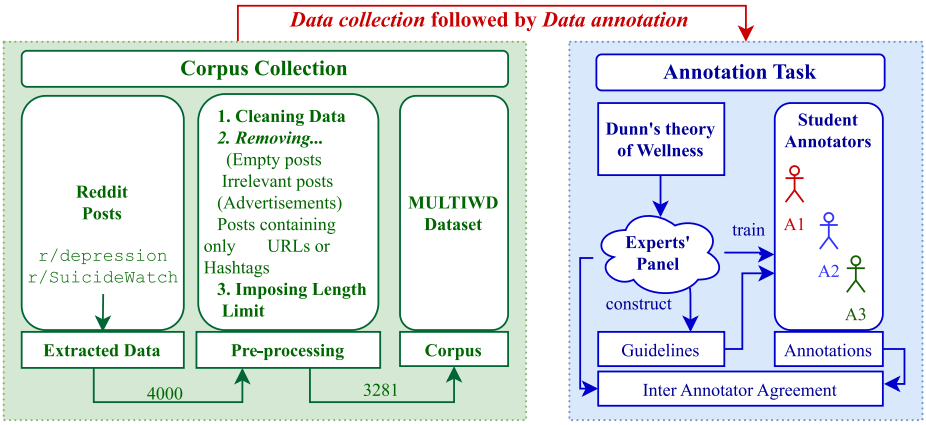


Fig. 2. Framework for Data Collection and Annotation Process. The diagram provides an overview of our approach, starting with the data collection and preprocessing stage, leading to the construction of a final corpus comprising 3281 samples. Trained annotators (A1, A2, A3) annotate the data under the supervision of our experts' panel, utilizing annotation guidelines prepared by them.

3.1.1. Data acquisition

There has been a notable rise in Reddit users discussing mental health issues in recent years [16]. This highlights the need to study online mental health discussions. We chose Reddit for our research because it allows open discussions on various mental health topics and lets users post anonymously, ensuring candid personal insights. We specifically extracted data from two popular subreddits: /depression and /SuicideWatch, as depicted in Fig. 2. While the Python Reddit API Wrapper (PRAW) API[3] provides an interface for collecting author's and posting information from Reddit. We consider ethical considerations while gathering and analyzing data from social media platforms to ensure privacy and transparency.

Search Query: We set the search query on Reddit, an optional anonymous social media platform in two online discussion communities: r/depression and r/SuicideWatch for English language posts.

Inclusion: We obtain an 4000 instances from Reddit through our search query between November 23, 2021, and January 4, 2022. An average of ≈100 data points per day is collected to ensure variation in the dataset.

Exclusion: We manually cleaned the data points. Upon initial screening, posts that are empty, irrelevant, or lack self-advocacy are identified and removed. We manually filter out empty posts, those with only URLs or social media handles, advertisements, and other irrelevant content. This ensures our dataset consists only of pertinent and significant data points.

We observe that users tend to write longer texts when sharing personal experiences, and lengthier comments usually get better responses compared to shorter ones [19]. While real-time Reddit posts can vary greatly in length, we standardized our dataset to a maximum of 300 words per post, resulting in a final collection of 3281 posts.

3.1.2. Annotation scheme

Our experts developed annotation guidelines that were based on the definitions of wellness dimensions provided by Dunn [3] by negotiating a trade-off between *text-based marking used for developing advanced AI models* and *reading between the lines to provide psychological insights*. With these annotation guidelines, we aim to achieve:

1. Identified wellness dimensions are accurate and align with Dunn's definitions.
2. The annotation process is consistent across different annotators with minimal errors and discrepancies.
3. Annotation process is efficient, allowing for the large amount of data to be annotated within a reasonable time frame.

The guidelines provide detailed instructions on how to identify and annotate text data for different wellness dimensions, using specific examples and criteria for each dimension. These guidelines also included instructions on how to resolve ambiguous cases, as well as guidance on how to ensure consistency and accuracy in the annotation process.

We employ three student interns from the postgraduate course and our experts train them through well-constructed annotation scheme. After training, the students were asked to annotate 40 samples as a group-activity. Clear instructions of were given to apply majority voting rule in-case of any discrepancies. Our expert panel evaluate the *sample set annotations*, and frame additional guidelines to ensure that the annotations are objective and consistent, which should lead to more reliable and trustworthy results. After three successful sample set annotations and guidelines up-gradation, we employ students to annotate the dataset as an individual-activity using fine-grained guidelines to avoid any potential biases that may affect the model's prediction. Based on the majority voting mechanism, our expert panel assign the final labels to the annotated dataset.

We obtain inter-annotator agreement among annotated dataset using Fleiss' Kappa coefficient.[4] To assess the inter-annotator agreement, the agreement on labels obtained for each of the six aspects is examined individually. We obtain the value of kappa coefficient (κ) for {SpA (58.34%), PA (68.66%), IA (75.29%), SA (83.14%), VA (76.77%), EA (64.23%)}. We obtain the average of agreement as $\kappa = 71.07\%$ which is considered to be a substantial agreement ([0.61–0.80]). In a sensitive, domain-specific, and psychology and emotion grounded task for multi-labeling, the inter-annotator agreement is often low [20] (see Table 1).

3.1.3. Data handling

In this section, we discuss the nature of MultiWD, including the imbalances in the number of labels across different aspects, limitations on Reddit post length, and adherence to FAIR principles for MultiWD.

Imbalanced dataset. We acknowledge the issue of imbalanced data distribution in MultiWD (see Table 2). Many posts in MultiWD are classified as Social Aspect, followed by Emotional Aspect and Physical Aspect, while other dimensions have significantly fewer instances. Such imbalanced dataset induce bias which developing machine learning models and affect their accuracy in predicting less frequent wellness

[3] https://praw.readthedocs.io/en/stable/.

[4] https://en.wikipedia.org/wiki/Fleiss%27_kappa.

Table 1
Dataset samples for MultiWD dataset.

Text	SpA	PA	IA	SA	VA	EA
It will all be over in one hour when society ends the season of expecting everyone to be fake joyous. FU society!	0	0	0	0	0	0
Yup. Everyone is out getting drunk and being reckless (like every other 19 year old). Here I am in a empty parking garage eating McDonald's because my parents forgot to save me some dinner. I hate my life.	0	0	0	1	0	0
I can't motivate myself to do anything at all, not even open up my laptop and do stuff with that... it is so awful to be stuck in that negative state of consciousness. :(what can I do? I take lyrica and milnacipran at the max dosage but it does not get better....	0	1	1	0	0	0
Too much. Over the past 4 years I've dealt with many suicides, overdoses and death. I'm tired and weak. It's hard to see where the light is at with ur eyes closed. God forgive me. I hope this post helps raise awareness. God bless	1	1	0	0	0	1
Winter semester begins tomorrow and I'm in denial. Back to group work, trying to make acquaintances, perform my best. I really just don't care, and thinking of going physically pains me.	0	0	1	1	1	1

Table 2
The statistics of MultiWD for different dimensions of wellness, indicating the imbalanced nature of dataset.

Dimensions	Count	Number of words				Number of sentences		
		Min/post	Avg/post	Max/post	Total	Avg/post	Max/post	Total
Spiritual Aspect	200	6	122.625	298	24 525	8.145	42	1629
Physical Aspect	923	8	134.678	300	124 308	8.55	32	7892
Intellectual Aspect	651	15	145.285	300	94 581	9.01	31	5869
Social Aspect	2129	1	132.728	300	282 580	8.29	42	17 652
Vocational Aspect	550	13	155.843	300	85 714	9.814	30	5398
Emotional Aspect	1661	1	131.414	300	218 280	8.279	31	13 752

dimensions. To address this issue, we suggest developing strategies for contextualized data augmentation methods at fine-grained level to balance out the dataset [21]. We observe the least number of instances in Spiritual Aspect having only 200 samples and the highest number of instances in Social Aspect having 2129 samples out of a total of 3281 samples in the dataset.

This skewed distribution of instances among different wellness dimensions exhibits an individual's tendency to express their *social experiences and emotions* more frequently than their *spiritual and intellectual experiences* on social media platforms. However, this observation does not imply that the Spiritual dimension is less important or less relevant in the context of mental health. It is plausible that some individuals may choose to communicate their spiritual experiences through alternative means of communication, such as in-person interactions or private messaging, rather than through public social media platforms like Reddit. Some may not be fully aware of the significance of spiritual experiences. Such instances incur relatively lower number of instances of the Spiritual Aspect in Reddit posts, as compared to other dimensions such as the Social Aspect.

Reddit post length. To ensure the compatibility of MultiWD with pre-trained models, we set the maximum length of each data sample to 300 words. We observe that the existing pre-trained models, such as BERT, is suggested to have a maximum input length of 512 tokens. Additionally, a smaller maximum length allows for faster training and inference times while still retaining sufficient contextual information in the text. We ensure that our dataset is of a manageable size for training and analysis. Thus, we limit the length of our samples to 300 word count.

3.2. FAIR principles

The FAIR (Find-able, Accessible, interoperable, and Reusable) principles are a set of guidelines designed to improve the quality and availability of research data. These principles were developed in response to the increasing volume of research data being generated and the need for better management and sharing of datasets.

1. According to the first principle, "Findable", we make MultiWD available in a publicly accessible repository of Github.[5]
2. The second principle, "Accessible" suggests data access and we make a potentially cleaner and the first version of MultiWD an open access ensuring its availability in a usable format of Comma-Separated Values (CSV).
3. The third principle, "Interoperable" advise the compatibility of MultiWD with other datasets, well assured with CSV format.
4. The fourth principle, "Reusable" suggest the use and reuse of dataset for different purposes, that is well aligned with its availability on Github.

3.3. Multi-label classification method

We present the challenge of labeling various wellness dimensions in user-written texts, highlighting the often-discussed areas of one's life with significant implications on mental health. Our primary focus is to effectively tag multiple wellness dimensions from user-penned texts. Given Dunn's theory's six dimensions, we leverage a multi-label classification approach using a deep NLP model.

Our text classification task, which involves six labels, aims to identify one or more predefined classes to which a given text belongs, based on corresponding text spans that assist in decision-making. However, these text spans do not specify named entities or any existing problem domain, underscoring the importance of detecting situations, perceptions and circumstances that indicate the impacted wellness dimension. Thus, we make decisions for text classification based on the text-spans, semantically indicating perception of the user.

3.3.1. Text representation

Given a user-penned text, let us denote it as t. We convert t into a high-dimensional vector space using embeddings. Other than using the feature vector representation and traditional embeddings such as BERT and Word2vec, we utilize an Open AI 1536-dimensional embedding model that is useful for tasks that require a high level of semantic understanding, as the large number of dimensions in the vector space allows for a finer-grained representation of the meaning of words and phrases. The resultant representation is:

$$T = \text{Embed}(t) \tag{1}$$

where Embed is an embedding function (for instance: BERT, Word2Vec, Open AI embedding).

[5] https://github.com/drmuskangarg/MultiWD.

The set of all wellness dimensions is defined as D:

$$D = \{D_1, D_2, \ldots, D_6\} \tag{2}$$

Here D_1: SpA, D_2: PA, D_3: IA, D_4: SA, D_5: VA, D_6: EA.

For each text T, the predicted output is a binary vector of 1 X 6 dimensions:

$$Y = [y_1, y_2, \ldots, y_6] \tag{3}$$

where y_i is 1 if T belongs to dimension D_i and 0 otherwise.

3.3.2. Architecture

We further introduce an architecture to align with the prospective solution of this task as M. Our model takes in a text representation T and outputs a vector in the space $[0,1]^6$, indicating the probability of the text belonging to each wellness dimension:

$$Y_{\text{predicted}} = M(T) \tag{4}$$

where each element of $Y_{\text{predicted}}$ is in the range [0,1], representing the probability of belonging to a particular dimension.

3.3.3. Loss function

To train M, we utilize a suitable loss function, say binary cross-entropy, given by:

$$L(Y_{\text{true}}, Y_{\text{predicted}}) = -\sum_{i=1}^{6} \left[Y_{\text{true},i} \log(Y_{\text{predicted},i}) + (1 - Y_{\text{true},i}) \log(1 - Y_{\text{predicted},i})\right] \tag{5}$$

where Y_{true} is the ground truth label vector for the text.

3.3.4. Objective

Our goal during training is to adjust the parameters of M to minimize the expected loss over our training data. Formally, given a dataset D of paired texts and labels, our objective is:

$$\min_M \mathbb{E}_{(T,Y_{\text{true}})\in D}[L(Y_{\text{true}}, M(T))] \tag{6}$$

Thus, the training of M is governed by a binary cross-entropy loss function. This function calculates the difference between the predicted values from the model and the actual ground truth. By leveraging an NLP model and optimizing it under a suitable loss function (tweak M's parameters to minimize this loss), we aim for accurate predictions that capture and classify wellness dimensions in a given text.

4. Experiments and evaluation

As a starting point of our study, we conduct thorough experiments by employing pre-trained language models, Open AI embeddings, and GPT-3 model to establish baseline.

4.1. Competing methods

The intuition behind using traditional learning-based models, large language models and domain-specific models is to establish the baseline for this novel task of identifying wellness dimensions. We first use the **Pre-trained Language Models** (PLMs) and **domain-specific Pre-trained Language Models** that allowed us to leverage the power of deep learning for analyzing MULTIWD. The PLMs that we use in this work are BERT [22], ALBERT [23], DistilBERT [24], and DeBERTa [25]. The domain-specific PLMs used for this study are PsychBERT [26], MentalBERT [27], and ClinicalBERT [28].

We further use **OpenAI embeddings API** to convert a given input text into *1536-dimensional* embedding - *text-embedding-ada-002* engine, a 1536-dimensional embedding model that is capable of capturing complex relationships between words and their meanings. The resulting embeddings are given as an input to train the learning-based traditional machine learning classifiers: Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), and XGBoost.

Our study evaluate the suitability of **LLMs** with N-shot learning in a sensitive domain of healthcare sector leveraging wellness dimensions. We further consider fine-tuning GPT-3.5-turbo and LLAMA to determine the effectiveness of Generative AI for domain-specific multi-label classification task in psychology-driven healthcare sector.

4.2. Experimental setup

Hyperparameter Settings: We split our data into 80:20 such that at least 20% of each of the wellness dimensions is present in the testing data, to ensure the integrity and reliability of models over unseen data. In Pre-trained language models, we first tokenize the input text using a PLMs tokenizer into a 768-dimensional vector, which is further given as an input vector to a fully connected network. We use a *binary cross-entropy* with *Log loss function* to train our models. We set max-length as 512 and other hyperparameters for our base model, BERT as batch-size 8 with early stopping at 20 epochs, Optimizer = Adam, learning rate = $2e{-}5$, weight decay = 0.01 and warm-up steps = 100 during training. In general, we train each model for 20 epochs, using a learning rate of $2e{-}5$ and a batch size of 8. We train Chatgpt models for four epochs, and kept all other parameters at their default values during the fine-tuning process. We set 8 samples for few-shot learning with GPT-3.5-turbo and LLAMA models.

Hardware Settings: All experiments were conducted in a Google Colaboratory (Colab) environment, a cloud-based platform that offers free access to GPUs for machine learning research.

4.3. Evaluation metrics and protocols

We use precision, recall, F1-score and accuracy to evaluate the performance of the models. Precision indicates the percentage of correctly identified instances that are truly relevant to a particular wellness dimension. The F1-score is prioritized over accuracy in our task because the goal is more about precisely identifying relevant instances for each wellness dimension, rather than achieving overall accurate classification of all instances.

The Matthews Correlation Coefficient (MCC) evaluates classification quality, providing scores from −1 (total disagreement) to 1 (perfect prediction), with 0 indicating random predictions. It is especially useful for imbalanced datasets like "MultiWD", where class sizes vary greatly. MCC can effectively assess a classifier's performance, even if it inaccurately predicts minority classes while still achieving high overall accuracy.

Next, Area Under the ROC Curve (AUROC) measures the ability of a classifier to distinguish between positive and negative examples by plotting the true positive rate (TPR) against the false positive rate (FPR) at various thresholds. The Area Under the Precision-Recall Curve (AUPRC) measures the trade-off between precision and recall across different classification thresholds. In our multi-label classification task involving multiple wellness dimensions per instance, accuracy alone is insufficient for assessing a model's performance in correctly predicting each label. Instead, AUROC and AUPRC offer a more detailed evaluation of the model's effectiveness across all labels.

4.4. Experimental results

We now discuss the experimental results that we present for **competing traditional classifers** with recorded precision, recall, F1-score, Accuracy and MCC as illustrated in Table 3. It displays the overall evaluation of the models where we discover the best performing models based on F1-score. We determine the efficacy of all models using MCC score. Next, we make the **aspect-based analysis** of different models through F1-score for fine-grained analysis. We discover the top performing models and suggest the need to balance-out dataset in the near future. Finally, we examine **AUROC and AUPRC curves** to analyze the overall quality of classifier.

Table 3
Experimental results with different classifiers. The bold values indicates the best performing models in corresponding set of classifiers.

Model	Precision	Recall	F1-score	Accuracy	MCC
BERT (*bert-base-uncased*)	73.74	81.38	**76.69**	**41.25**	**66.60**
ALBERT (*albert-base-v2*)	74.11	75.12	74.26	37.44	63.95
DistilBERT (*distilbert-base-uncased*)	72.95	78.67	75.43	36.38	64.97
DeBERTa (*deberta-base*)	63.38	91.02	74.45	23.44	61.74
ClinicalBERT (*Bio-ClinicalBERT*)	70.86	77.10	73.41	34.25	62.08
MentalBERT (*mental-bert-base-uncased*)	72.88	80.48	**76.19**	**37.14**	**65.47**
PsychBERT (*psychbert-cased*)	71.87	76.69	73.92	34.86	62.60
OpenAI + LR	68.39	55.03	53.34	26.79	50.08
OpenAI + SVM	76.90	58.72	61.98	31.02	55.45
OpenAI + RF	76.35	52.80	53.42	25.19	48.74
OpenAI + MLP	69.11	66.49	**67.37**	**28.66**	**55.72**
OpenAI + XGBoost	73.64	60.39	63.63	28.24	53.76
Zero-shot learning GPT-3.5-turbo (20B)	29.78	60.54	38.44	57.58	26.66
Zero-shot learning LLAMA (70B)	59.46	57.19	52.71	74.25	28.30
One-shot learning GPT-3.5-turbo (20B)	32.07	62.42	36.16	49.44	29.06
One-shot learning LLAMA (70B)	49.25	54.33	48.87	71.74	24.24
Few-shot learning GPT-3.5-turbo (20B)	54.94	64.08	**56.26**	**74.38**	**45.36**
Few-shot learning LLAMA (70B)	49.01	62.42	49.04	63.75	25.71
Fine-tuned GPT-3 (200B)	75.13	76.94	**75.94**	41.55	**65.19**
Fine-tuned GPT-3.5-turbo model (20B)	69.53	70.53	69.93	**85.06**	56.22
Fine-tuned LLAMA (13B)	34.12	26.67	28.94	66.03	03.36

4.4.1. Competing methods

The F-measure penalizes models for inadequate performance on the minority class, which is an essential aspect of model evaluation. We illustrate the results in Table 3. Out of all the PLMs, the BERT (bert-base-uncased) model exhibits the highest F1-score of 76.69%, demonstrating its efficacy in capturing the intricate relationships between the input features and the output labels. MentalBERT, among the Contextual-PLMs, achieves the highest performance and is on par with the BERT-base model, indicating a strong correlation between wellness dimensions and mental health in social media texts. The MLP model utilizing OpenAI embeddings surpassed other learning-based models, achieving an F1-score of 67.37%. However, it is not as efficient as PLMs. The results of n-shot learning demonstrate that few-shot learning outperformed other n-shot learning methods, achieving an F1-score of 63.70%. However, it exhibited poorer performance compared to both PLMs and learning-based models. Our findings indicate that few-shot learning is not a viable approach for enhancing the performance of a sensitive domain-specific NLP task, such as healthcare, particularly in the context of mental health analysis.

We observe the N-shot learning models over LLMs pre-trained over huge amount of data with 20B parameters in GPT-3.5-turbo and 70B parameters in LLAMA model. We observe the best performance with Few-shot learning GPT-3.5-turbo, as observed from the F1-score and accuracy scores. The fine-grained analysis with different wellness dimensions show increased consistency and high performance with SA and EA. We compared three Large Language Models (LLMs), GPT-3 with 200 billion parameters, ChatGPT with 20 billion parameters and LLAMA with 13 billion parameters, and observed that ChatGPT outperforms LLAMA models. The experimental results with LLMs vary for SPA and VA, likely because psychological inferences require more human intervention. We acknowledge that due to resource availability, we have trained LLAMA model on 240 samples which could be the reason of poor performance other than less number of parameters. Across all models, the accuracy scores were comparatively lower when compared to the F1 score. This disparity can be attributed to the evaluation criteria employed, which considers a prediction as accurate only if all six multiple label predictions are correct. Any incorrect prediction for even a single label leads to a zero accuracy score.

4.4.2. Fine-grained aspect based analysis

We further illustrate the performance of all classifiers for each wellness dimension in Fig. 3. Spiritual Aspect gives the worst performance and social aspect gives the best performance, signifying the importance of more number of samples, required for identifying wellness dimensions. Our experiments showed that the performance of the N-shot learning models was consistently lower as compared to other machine learning models, suggesting that direct application of GPT models are not well-suited for our task. The mental health-related text is often complex and nuanced, requiring a deeper understanding of underlying context and language. GPT models, while effective in many NLP-centered tasks, are not sufficiently equipped to capture these nuances and contextual factors for wellness dimensions. We further observe comparable performance with PA, EA, IA and VA. Interestingly, the machine learning models leveraging Open AI embeddings show poor performance for IA and VA, suggesting the inadequacy of generalized representations offered by Open AI embeddings deployed to capture Wellness dimensions.

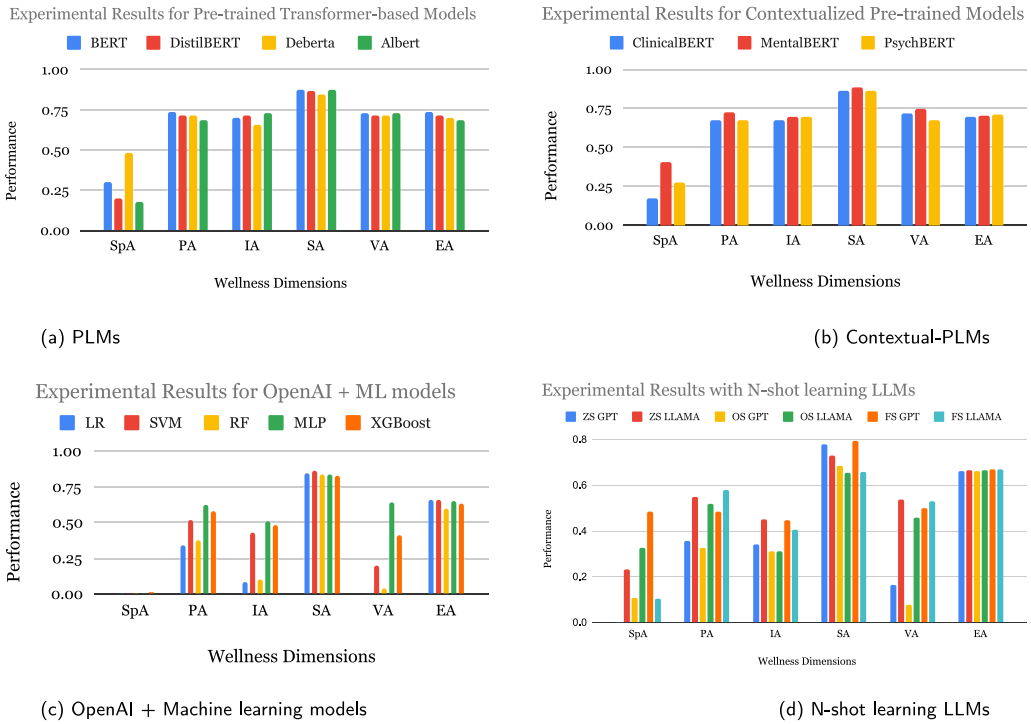
4.4.3. AUROC and AUPRC curves

The Fig. 4 presents the AUC and ROC curves for the (a) PLMs and (b) Contextual-PLMs. The AUROC analysis show that both the BERT and MentalBERT models have the highest coverage of the maximum area. The strong performance of these models is particularly noteworthy, as it indicates that they are capable of accurately identifying instances that are truly positive or negative, while minimizing the occurrence of false positives and false negatives. This high level of accuracy is a critical factor for applications of this nature, where precise identification of wellness dimensions is necessary for effective decision-making and interventions. In our analysis of the AUPRC curves, we found that the DistilBERT and MentalBERT models provided the highest coverage of the maximum area. We noticed that DistilBERT achieved a higher area under the curve compared to other models, but most of the gain occurred for lower recall values. This indicates that DistilBERT has a higher precision but struggles with recall, implying that it may miss out on important instances of the target label. In contrast, BERT had a more balanced trade-off between precision and recall, making it a better choice for the multi-label classification task of identifying wellness dimensions in social media posts.

5. Discussion

5.1. Semantic word ambiguity

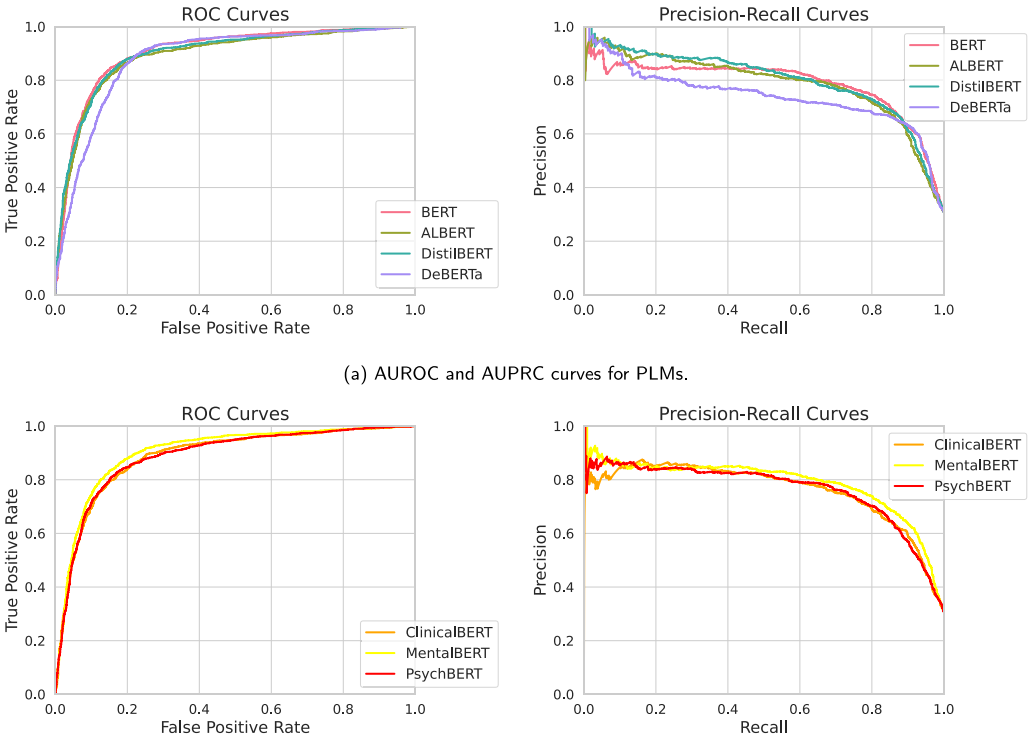
A major issue of semantic word ambiguity arises as it analyze the text data in the context of psychological concepts, which can lead to multiple interpretations of the same word or phrase. For instance, consider the posts P_2 and P_3 given as examples.



(a) PLMs (b) Contextual-PLMs

(c) OpenAI + Machine learning models (d) N-shot learning LLMs

Fig. 3. Experimental results for fine-grained performance evaluation and aspect-based analysis through existing classifiers.



(a) AUROC and AUPRC curves for PLMs.

(b) AUROC and AUPRC curves for Contextual-PLMs.

Fig. 4. AUROC and AUPRC for PLMs and Contextual-PLMs.

P_2: ...bad luck due to demons in my head...

P_3: ...head-ache or injury in my head...

The word "head" is used in both posts, but they have different interpretations based on the context. Post P_2 discusses bad luck due to demons in the head, which is related to the psychological concept of supernatural beliefs, and therefore affects the SpA. On the other hand, post P_3 pertains to physical injury or incapability causing mental disturbance, which is related to the PA dimension.

5.2. Metaphors

The use of metaphors is a common cultural practice in social media engagement, and it poses a significant challenge in accurately analyzing the data for wellness dimensions. Consider the following posts P_1 and P_2:

P_1: ...maybe I'll drink myself to death before I wind up homeless...

P_2: ...I drink a lot of alcohol...

Both the posts P_1 and P_2 contain the word "drink". However, the meaning of the word is different in both posts due to the use of metaphors. In P_1, the expression "drink myself to death" is a metaphor that suggests the risk of homelessness, indicating VA. In contrast, in P_2, the word "drink" refers to the consumption of alcohol, indicating physical unfitness or PA. Metaphors are culture specific and it is essential to incorporate cultural understanding while developing AI models for this task.

5.3. Attention and ambiguity

Attention refers to the ability to identify words or phrases that are more important in a post, while ambiguity pertains to the multiple possible interpretations of a word or phrase. To reduce ambiguity, low-level analysis and natural language understanding are used to identify the most important words in a post, even if those same words are less important in other posts.

P_0: From dealing with the fallout of my ex, to stressors at work, nothing compares to true loss of my baby boy. To everyone feeling shitty this New Year's Eve, you are not alone.

P_1: My mom says I cant work and controls my life...

P_2: ...soul sucking job...

P_3: ...unable to connect with my soul...

P_4: I have 0 friends who would talk to me outside of work...

For instance, in the given example, post P_0 contains multiple words that reflect different wellness dimensions, such as *feeling shitty* and *work* reflecting EA and VA, respectively. However, the most important phrase in this post is *loss of my baby boy*, reflecting the SA category. Similarly, in posts P_1 and P_4, the words *work* and *friends* must be emphasized to assign them to the appropriate wellness dimensions. The word *soul* in posts P_2 and P_3 must also be analyzed carefully to identify whether it is used as an adjective or a noun, thus determining whether it reflects VA or SpA, respectively.

5.4. Informal text

Social media users often use informal language and the unstructured nature of social media text can make it difficult to accurately incorporate external knowledge. This unstructured nature of data results in ambiguity, incorrect interpretations, or inconsistencies in the meaning of a given text, which can negatively impact the effectiveness of models.

Limitations

Because annotation is a subjective process, we anticipate that our gold-labeled data and the distribution of labels in our dataset may contain some biases. As the inter-annotator agreement (κ-score) was found to be high, we have confidence that the annotation instructions were correctly assigned for the majority of the data points. We acknowledge that this study includes the user-penned text that are available on Reddit platform only. As Reddit users have the option of posting anonymously, the posts may contain irrelevant and somatic syndrome disorder, depicting psychological distress in the form of physical issues impacting multiple well-being dimensions. However, our study set the starting point of tagging wellness dimensions in a given text.

Ethics and broader impact

Social media data is often personal and sensitive in nature. The dataset used in this study was collected from Reddit, an online platform where users can post anonymously and their IDs are kept anonymous. Additionally, to safeguard user privacy and prevent misuse, all sample posts showcased in this research have been anonymized, obfuscated, and rephrased. In order to comply with privacy regulations, we have taken care not to reveal any personal information, including demographic information, location, and personal details of social media users, while making the dataset available. Furthermore, we have made our dataset available on Github, enabling the baseline results to be easily reproduced.

6. Conclusion

Our work is the first of its kind that emphasizes the necessity of identifying and classifying wellness dimensions for mental health analysis in Reddit posts. We first develop annotation scheme to discover the presence of one of the six well-established wellness dimensions introduced by Dunn. Next, we perform the annotation task to construct and release an annotated dataset, MultiWD, of 3281 user-penned text. Through our experiments with competing methods, we demonstrate the potential and emphasize the need to improve the accuracy and comprehensiveness of AI models for this task. Based on our findings, we introduce BERT and MentalBERT as the baseline methods with F1-score as 76.69% and 76.19%, respectively. Interestingly, our study reveals the need of domain-specific and/or specialized large language models for psychology-driven healthcare utilization. In future, we plan to infuse external knowledge to develop a more comprehensive AI model for this task. We also plan to perform experiments with GPT4, advanced prompt engineering methods, prompt optimization, chain-of-thought and chain-of-knowledge, in future, to observe reliability and trustworthiness of LLMs for this task.

CRediT authorship contribution statement

Muskan Garg: Conceptualization, Data curation, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. **Xingyi Liu:** Investigation, Methodology, Resources. **M.S.V.P.J. Sathvik:** Methodology. **Shaina Raza:** Writing – review & editing. **Sunghwan Sohn:** Supervision, Funding acquisition, Project administration, Writing – review & editing, Resources.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Muskan Garg reports financial support was provided by Mayo Clinic Minnesota. Muskan Garg reports a relationship with Mayo Clinic Minnesota that includes: employment. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Publicly available on Github.

Acknowledgments

This project was supported by NIH R01 AG068007. We extend our sincere acknowledgment to the postgraduate student annotators, Ritika Bhardwaj, Astha Jain, and Amrit Chadha, for their diligent efforts in the annotation process. We express our gratitude to Veena Krishnan, a senior clinical psychologist, and Ruchi Joshi, a rehabilitation counselor, for their unwavering support throughout the project.

References

[1] UN, Transforming our world: The agenda 2030 for sustainable development, 2015.
[2] U. ESCAP, W. WHO, et al., SDG 3 Goodhealth and Well-Being: Ensure Healthy Lives and Promote Well-Being for All at All Ages, United Nations, 2021.
[3] H.L. Dunn, High-level wellness for man and society, Am. J. Public Health Nations Health 49 (6) (1959) 786–792.
[4] B. Pan, H. Wu, X. Zhang, The effect of trait mindfulness on subjective well-being of kindergarten teachers: The sequential mediating roles of emotional intelligence and work–family balance, Psychol. Res. Behav. Manage. (2022) 2815–2830.
[5] A.G. Abraham, C. Hong, J.A. Deal, B.M. Bettcher, V.S. Pelak, A. Gross, K. Jiang, B. Swenor, W. Wittich, Are cognitive researchers ignoring their senses? The problem of sensory deficit in cognitive aging research, J. Amer. Geriatr. Soc. (2023).
[6] G. Coppersmith, M. Dredze, C. Harman, Quantifying mental health signals in Twitter, in: Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, 2014, pp. 51–60.
[7] U. Pavalanathan, J. Eisenstein, Emoticons vs. emojis on Twitter: A causal inference approach, 2015, arXiv preprint arXiv:1510.08480.
[8] A. Zirikly, M. Dredze, Explaining models of mental health via clinically grounded auxiliary tasks, in: Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology, 2022, pp. 30–39.
[9] G. Ansari, M. Garg, C. Saxena, Data augmentation for mental health classification on social media, 2021, arXiv preprint arXiv:2112.10064.
[10] S. Ghosh, S. Roy, A. Ekbal, P. Bhattacharyya, CARES: Cause recognition for emotion in suicide notes, in: European Conference on Information Retrieval, Springer, 2022, pp. 128–136.
[11] M. Garg, C. Saxena, S. Saha, V. Krishnan, R. Joshi, V. Mago, CAMS: An annotated corpus for causal analysis of mental health issues in social media posts, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 6387–6396.
[12] Y. Dang, F. Li, X. Hu, V.K. Keloth, M. Zhang, S. Fu, M.F. Amith, J.W. Fan, J. Du, E. Yu, et al., Systematic design and data-driven evaluation of social determinants of health ontology (SDoHO), J. Am. Med. Inform. Assoc. (2023) ocad096.
[13] B.G. Patra, M.M. Sharma, V. Vekaria, P. Adekkanattu, O.V. Patterson, B. Glicksberg, L.A. Lepow, E. Ryu, J.M. Biernacka, A. Furmanchuk, et al., Extracting social determinants of health from electronic health records using natural language processing: a systematic review, J. Am. Med. Inform. Assoc. 28 (12) (2021) 2716–2727.
[14] M. Garg, Mental disturbance impacting wellness dimensions: Resources and open research directions, Asian J. Psychiatry (2023) 103876.
[15] M. Garg, WellXplain: Wellness concept extraction and classification in Reddit posts for mental health analysis, Knowl.-Based Syst. (2023) 111228.
[16] M. Garg, Mental health analysis in social media posts: A survey, Arch. Comput. Methods Eng. (2023) 1–24.
[17] P. Kumar, M. Vardhan, PWEBSA: Twitter sentiment analysis by combining plutchik wheel of emotion and word embedding, Int. J. Inf. Technol. (2022) 1–9.
[18] X. Qu, J. Zeng, D. Liu, Z. Wang, B. Huai, P. Zhou, Distantly-supervised named entity recognition with adaptive teacher learning and fine-grained student ensemble, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, 2023, pp. 13501–13509.
[19] S. Park, I. Kim, S.W. Lee, J. Yoo, B. Jeong, M. Cha, Manifestation of depression and loneliness on social networks: a case study of young adults on facebook, in: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, 2015, pp. 557–570.
[20] G. Roccabruna, S. Azzolin, G. Riccardi, Multi-source multi-domain sentiment analysis with BERT-based models, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 581–589.
[21] A.N. Tarekegn, M. Giacobini, K. Michalak, A review of methods for imbalanced multi-label classification, Pattern Recognit. 118 (2021) 107965.
[22] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, 2019, arXiv:1810.04805.
[23] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, ALBERT: A lite BERT for self-supervised learning of language representations, 2020, arXiv:1909.11942.
[24] V. Sanh, L. Debut, J. Chaumond, T. Wolf, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, 2020, arXiv:1910.01108.
[25] P. He, X. Liu, J. Gao, W. Chen, DeBERTa: Decoding-enhanced BERT with disentangled attention, 2021, arXiv:2006.03654.
[26] V. Vajre, M. Naylor, U. Kamath, A. Shehu, PsychBERT: A mental health language model for social media mental health behavioral analysis, in: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2021, pp. 1077–1082, http://dx.doi.org/10.1109/BIBM52615.2021.9669469.
[27] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, E. Cambria, MentalBERT: Publicly available pretrained language models for mental healthcare, 2021, arXiv:2110.15621.
[28] K. Huang, J. Altosaar, R. Ranganath, ClinicalBERT: Modeling clinical notes and predicting hospital readmission, 2020, arXiv:1904.05342.