MULTIWD: Multiple Wellness Dimensions in Social Media Posts

DOI

10.36227/techrxiv.22816586.v1

MULTIWD: Multiple Wellness Dimensions in Social Media Posts

MSVPJ Sathvik, *Student Member, IEEE,* Muskan Garg

***Abstract*—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. In our research, we introduce a fine-grained task focused on identifying indicators of wellness dimensions and mark their presence in self-narrated human-writings on Reddit social media platform. Our work serves as a valuable ground work for the early detection of chronic mental disturbances. 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 utilize existing classifiers to solve this multi-label classification task and introduce them as baselines. We highlight practical implications and our findings indicate the potential for further enhancements to develop more comprehensive and contextually-aware AI models in this domain.**

***Index Terms*—dataset, mental health, multi-label classification, social media, wellness dimensions**

I. INTRODUCTION

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] Mental disturbance is caused by genetics, environment, and other life experiences. Mental disorders, have significant impact on an individual's life, 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.

A. Problem Formulation

As a starting point of our study, we approach the problem of identifying multiple wellness dimensions discussed in a Reddit post 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, by considering the multiple wellness dimensions in this post, we can develop AI models that are better equipped to identify and emphasise particular aspects of deteriorating mental health. Such comprehensive and contextual AI models detect the affected wellness dimensions at an early stage and provide appropriate interventions to prevent the progression of mental illness into a chronic state.

MSVPJ Sathvik is with Indian Institute of Information Technology Dharwad, Hubli city, Karnataka, India. E-mail: 20bec024@iiitdwd.ac.in

Muskan Garg is with Mayo Clinic, Rochester, USA. Email:garg.muskan@mayo.edu, muskanphd@gmail.com

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

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

B. Applications

1) Interpersonal Risk Factors: According to a 2019 survey conducted by the Substance Abuse and Mental Health Services Administration (SAMHSA), individuals without strong familial and friendship ties are ten times more prone to experiencing mental health challenges as compared to those with a robust support system [6]. Moreover, people who have experienced adverse life events, such as trauma, abuse, or neglect, are more likely to struggle with mental health issues if they lack adequate social support. The Interpersonal-Psychological Theory of Suicide, as described by [7], argues that the development of serious suicidal behavior results from a combination of three proximal factors: *(i) acquired capability, (ii) perceived burdensomeness (PBu), and (iii) thwarted belongingness (TBe)*,[3] which are necessary and sufficient causes of suicidal behavior. Thus, the Interpersonal-Psychological Theory of Suicide highlights the importance of considering wellness dimensions in the development and prevention of suicidal behavior.

2) Supporting Cohort Studies: Longitudinal studies ensure fair and accountable practices of examining changes in users' emotions over time. Our dataset can potentially benefit ongoing research aimed at identifying these changes from users' historical timeline through social media posts [8]. Consider the following set of posts posted by a user A for varying time intervals $t = \{t1, t2, t3, t4, t5\}$:

> *T1:* ...I am not entirely sure; I am making sense of my life...
> *T2:* Politics is disastrous, and I am in the middle.
> *T3:* It drives me on the edge and restless to see what the outcome would be.
> *T4:* I hope there would be a life without politics. My relationship with my wife is also political.
> *T5:* I need better life after my death.

The posts range from expressing confusion (*T1*) to discussing politics (*T2-T4*) and finally to expressing a desire for a better life after death (*T5*). Such progression illustrates the significant impact that users' experiences have on different wellness dimensions over the period of time. Our dataset has the potential to complement and support other evolving studies such as [9] as an intrinsic multi-label classification task.

3) Pre-screening and Triaging: Pre-screening is a preliminary evaluation process that helps to identify individuals who may require further assessment or treatment [10]. *Mental health triaging* is a process of assessing the level of urgency and severity of a person's mental health needs and determining the appropriate course of action or treatment [11]. The identification of wellness dimensions plays a vital role in prescreening and triaging of mental health (see Figure 1). Our work facilitates the prioritization of individuals who require immediate attention and direct them to the appropriate mental health services with appropriate intervention of mental health practitioners and clinical psychologists.

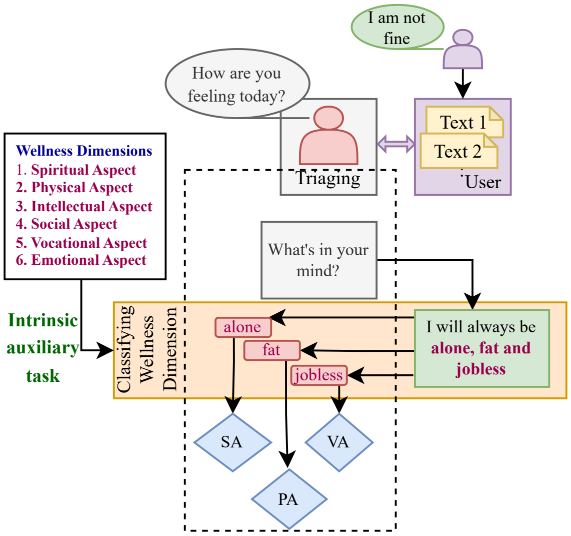


Fig. 1. Overview of the Mental health triaging and pre-screening application. Wellness dimensions as an intrinsic auxiliary task. Figure represents three different aspects present in one single sentence. Here SA: Social Aspect, PA: Physical Aspect, VA: Vocational Aspect.

C. 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 aids the early detection of chronic mental disturbance.
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.

II. Background

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. An individual with physical symptoms related to mental disturbance might require interventions related to exercise and nutrition.

A. Simulation of In-person Therapy Sessions

In face-to-face therapy sessions, therapists rely on manual techniques to identify the underlying causes and consequences of mental disturbances that may trigger latent factors, leading to severe chronic mental health disorders and cognitive decline. To handle this, they ask questions, observe the patient's

[3] Acquired capability refers to a desensitization to pain and fear of death, which can be developed through repeated exposure to painful or frightening experiences. PBu is the feeling of being a burden on others or society, while TBe is the feeling of being disconnected or isolated from others.

behavior, and analyze their thought patterns to gain insight into the root causes of their mental health disturbance [12].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.

B. In-person Sessions and Social Media Data

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. Moreover, we harness the social media data at a much larger scale than traditional in-person sessions. Social media enables real-time posting, which allows for the tracking of changes in wellness dimensions over the period of time facilitating the identification of patterns and trends that may not be readily apparent in in-person sessions. However, there are concerns about the ethical implications of using such data to protect the privacy and well-being of individuals. Moreover, AI models must be trained in an unbiased and ethical manner.

C. Mental Health Analysis in Social Media Posts

While progress has been made in developing classifiers and gathering longitudinal social media data, there is a lack of research and developments in fine-grained analysis. For instance, [13] 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 [14]–[16]. [17] used a multimodal deep learning approach to diagnose depression, anxiety, and stress.

While previous studies have utilized computational intelligence to identify mental health issues in social media data, they have primarily focused on specific mental health conditions such as depression, anxiety, and suicidal behavior. Mental health is a multi-dimensional construct that encompasses various aspects leveraging wellness dimensions. Thus, the inclusion of wellness dimensions in mental health analysis can provide a more comprehensive understanding of an individual’s mental well-being. Therefore, our research aims to address this gap in the literature by proposing a fine-grained problem that considers multiple dimensions of wellness to better understand mental health in social media data.

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 [12], [27], fostering the need of explicitly considering wellness dimensions as an intrinsic fine-grained task (see Table I).

III. High-level Wellness theory

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]. 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:

a) Spiritual Aspect (SpA).: A wellness dimension concerned with the search for meaning and the purpose in human existence. It encompasses the development of a deep appreciation for the vastness and complexity of life, and for the natural forces that exist in the universe, thereby cultivating a sense of inner peace, tranquility, and mindfulness. People connect with meditation, prayer, yoga, spending time in nature, or participating in religious or cultural ceremonies.

b) Physical Aspect (PA).: The process of physical development involves not only the growth and changes in the body, but also the adoption of healthy habits that can promote well-being. A key component of this is maintaining a good diet and nutrition, as well as avoiding harmful substances such as tobacco, drugs, and excessive alcohol consumption. When people are subjected to body shaming, they may become more aware of their physical appearance and any related medical history, leading to feelings of shame, self-doubt, and even depression.

c) Intellectual Aspect (IA).: A wellness dimensions that supports incorporating intellectual and cultural activities both inside and outside of the classroom, and taking advantage of the available human and learning resources. People enhance their cognitive abilities to gain a deeper understanding of the world around them. IA based activities include reading, attending lectures, participating in discussions and debates, pursuing creative hobbies, education and engaging in cultural events and experiences.

d) Social Aspect (SA).: The interdependence between individuals, society, and the natural world, recognize one’s place in society and the impact that they have on environment. Through social connections, individuals can flourish in their interpersonal relationships, enhancing their ability to empathize with and appreciate cultural differences. Building and maintaining strong social connections with family, friends, coworkers, and community members can promote feelings of belonging and support.

e) Vocational Aspect (VA).: The work can bring a sense of satisfaction and enrichment to one’s life, affecting an individual’s attitudes towards creativity, professional growth, and financial management. Occupational fulfillment contributes positively to their overall sense of happiness and purpose,

TABLE I
A CORPUS COLLECTION AND THEIR AVAILABILITY FOR MENTAL HEALTHCARE. A: AVAILABLE, ASA: AVAILABLE VIA SIGNED AGREEMENT, AR: AVAILABLE ON REQUEST FOR RESEARCH WORK

Dataset	**Task**	**Avail.**
CLPsych [18]	Depression detection for suicide risk	S
MDDL [19]	Depression candidate detection (D1, D2, D3)	A
RSDD [20]	Depression detection from Reddit data	ASA
SMHD [21]	Multi-task mental illness from Reddit data	ASA
eRISK [22]	Early risk detection: CLEF	A
Sina Weibo [23]	Identifying candidates with suicide risk	AR
Dreaddit [24]	Stress detection from Reddit posts	A
UMD-RD [25]	Suicide risk detection from Reddit data	ASA
SDCNL [26]	Suicide v/s depression from Reddit	A
Ghosh *et. al.,* [27]	Cause detection and explanation in Suicide Notes	AR
CAMS [12]	Interpretable Causal analysis from Reddit	A
MULTIWD (Ours)	Multi-label Wellness Dimensions in Reddit Post	A

channelizing the motivation, productivity, and creativity in the workplace. The occupational dimension also recognizes the potential negative impact of work-related stress, burnout, and financial strain. When work becomes a source of anxiety or dissatisfaction, it can have detrimental effects on an individual's mind.

f) Emotional Aspect (EA).: The emotional dimension of wellness is concerned with enhancing an individual's awareness and acceptance of their own feelings. We use a set of 13 well-defined labels were assigned to emotions based on the Plutchik's wheel framework. These labels were then organized into pairs of polar opposites, resulting in four binary classification tasks: Love vs. Hate, Joy vs. Sadness, Trust vs. Disgust, and Anticipation vs. Surprise [28].

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.

IV. CORPUS CONSTRUCTION

In our research, we aim to accurately identify wellness dimensions in Reddit data, which can be a challenging and error-prone task due to the subjectivity and complexity of the concept. A naive approach to label without considering this complexity may lead to errors and inaccuracies in the dataset. Therefore, we recognize the importance of establishing a well-defined and consistent annotation process to ensure the accuracy and reliability of our dataset. To this end, we establish an expert panel consisting of a clinical psychologist, a rehabilitation counsellor, and a social NLP researcher. This diverse team was able to provide a wide range of perspectives and expertise to the annotation process. The annotation guidelines were established in a collaborative manner, with an intense discussion by team members for refining the guidelines, and facilitate accurate and consistent annotation scheme.

A. Data Acquisition

In recent years, there has been a significant increase in the number of Reddit subscriber who discuss mental health-related issues [29]. This trend underscores the importance of studying online discussions related to mental disturbance. Thus, we use Reddit, a widely-used social media platform that offers users an open forum to participate in discussions on a diverse array of mental health subjects. Its distinctive attribute of permitting individuals to post anonymously renders it a well-suited platform for amassing candid and personal data as first-hand experiences. We extract data from two of the most popular subreddits: `/depression` and `/SuicideWatch` as shown in Figure 2.

While the Python Reddit API Wrapper (PRAW) API[4] provides an interface for collecting data from Reddit, it is important to consider ethical considerations when gathering and analyzing data from social media platforms. Although PRAW allows author's and posting information retrieval, we extract only post's title and text to ensure the privacy of social media users' posts. We obtain an 4000 instances from Reddit on r/depression and r/SuicideWatch between November 23, 2021, and January 4, 2022. An average of ≈ 100 data points per day is collected to ensure variation in the dataset. We manually cleaned the data points.

Posts that lack self-advocacy or are empty/irrelevant are identified upon initial screening. Manual screening and filtering are performed by removing all the empty posts, the posts containing only URLs or other social media handles, advertisements and irrelevant data. We finally ensure that our dataset only contains relevant and meaningful data points. We have found that individuals tend to write more extended texts when they share personal experiences, which aligns with conventional arguments that prolonged remarks gather better responses from others in comparison to transient remarks [30]. The length of real-time Reddit posts ranges from a few characters to thousands of words. To standardize our data, we set a maximum length of 300 words for each post, resulting in a final corpus of 3281 posts.

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

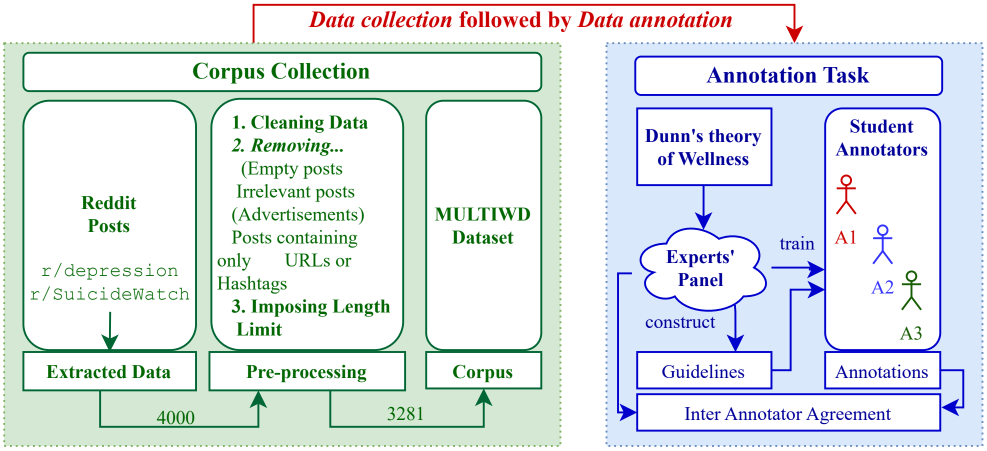


Fig. 2. Framework for Data Collection and Annotation Process. The diagram provides an overview of our approach, starting with the data collection and preprocessing stage, leading to the construction of a final corpus comprising 3281 samples. Trained annotators (A1, A2, A3) annotate the data under the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

B. 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.

TABLE II
INTERPRETATION OF RESULTING VALUES OF FLIESS' KAPPA AGREEMENT STUDY [32].

Range	Interpretation
< 0:	Less than chance agreement
0.01–0.20:	Slight agreement
0.21– 0.40:	Fair agreement
0.41–0.60:	Moderate agreement
0.61–0.80:	Substantial agreement
0.81–0.99:	Almost perfect agreement

We obtain inter-annotator agreement among annotated dataset using Fleiss' Kappa coefficient[5]. 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 for {SPA, PA, IA, SA, VA, EA} as $\kappa = \{58.34\%, 68.66\%, 75.29\%, 83.14\%, 76.77\%, 64.23\%\}$, respectively. We obtain the average of agreement as $\kappa = 71.07\%$ which is considered to be a substantial agreement (see Table II). In a sensitive, domain-specific, and psychology and emotion grounded task for multi-labeling, the inter-annotator agreement is often low [31].

C. Discussion

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.

1) Imbalanced dataset: We acknowledge the issue of imbalanced data distribution in MULTIWD (see Table IV). 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 dimensions. To address this issue, we suggest developing strategies for contextualized data augmentation methods at

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

TABLE III
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
I don't know what is happening to my mind. I have got 3 panic attacks in the last 5 days. Everything is getting messed up. Things are going downhill at my home. I can't sleep at night now for the shoutings and all the fights. I can't study even though exams are coming. My soul love just got engaged to someone else and I have to see tgose the moment I step out of my house as he is my neighbour. I always wanted to go abroad and make a living as a nurse. I am falling in my grades and all my former classmates are settling down overseas. My relatives hate me because I am bad at studies and a fat and ugly girl. I can't take this anymore. And also my depression is having a great time feeding over my faliures!!	0	1	1	1	1	1

fine-grained level to balance out the dataset [33]. 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.

2) 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) FAIR Principles: The FAIR (Findable, 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) The first principle, "Findable" ensures that the data and metadata should be easy to find and identify. We make MULTIWD findable by depositing the dataset in a publicly accessible repository of Github.[6]
2) The second principle, "Accessible" suggests that data should be available to anyone who wants to access it. We have made a potentially cleaner and the first version of MULTIWD, accessible by providing open access to it and ensuring that it is available in a usable format of Comma-Separated Values (CSV).
3) The third principle, "Interoperable" advise that data should be structured in a way that allows it to be easily combined with other datasets. To ensure ethical integrity, we do not disclose the metadata of users or posting behaviour, instead we make our dataset interoperable by structuring it in the six different columns of CSV file which is compatible with existing mental health datasets.
4) The fourth principle, "Reusable" means that data should be designed and structured in a way that allows it to be used and reused for different purposes. We discuss the annotation scheme used to construct the MULTIWD dataset by a panel of experts and the annotation task in this research paper and on Github Repository.

V. EXPERIMENTS AND EVALUATION

We further simulate the experimentation with traditional classifiers on MULTIWD. In this section, we begin with briefing the experimental setup and performance evaluation measures. Next, we elaborate the experimental results and analyze errors. Finally, we discuss ethical considerations and broader impact.

A. Experimental Setup

For our experiments, we employ three types of conventional classifiers.

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

TABLE IV
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	24525	8.145	42	1629
Physical Aspect	923	8	134.678	300	124308	8.55	32	7892
Intellectual Aspect	651	15	145.285	300	94581	9.01	31	5869
Social Aspect	2129	1	132.728	300	282580	8.29	42	17652
Vocational Aspect	550	13	155.843	300	85714	9.814	30	5398
Emotional Aspect	1661	1	131.414	300	218280	8.279	31	13752

1) Pre-trained Language Models (PLMs): We first use the **Pre-trained Language Models** (PLM) and **Contextualised Pre-trained Language Models (Contextual-PLMs)** that allowed us to leverage the power of deep learning for analyzing MULTIWD. The PLMs that we use in this work are BERT [34], ALBERT [35], DistilBERT [36], and DeBERTa [37]. The Contextual-PLMs used for this study are PsychBERT [38], MentalBERT [39], and ClinicalBERT [40]. 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 train each model for a total of 20 epochs, using a learning rate of $2e-5$ and a batch size of 8.

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

3) GPT models - N-shot learning and fine-tuning: We finally explore the effectiveness of Open AI in classifying input data using **zero-shot, one-shot, and few-shot learning** approaches. Zero-shot learning is performed by providing prompts used to make predictions about new instances without training the model. One-shot learning involves training a model with only one example of each class where a given model learns to recognize a new class based on a single example by generalizing from the examples it has seen in the past. Few-shot learning is similar to one-shot learning but involves providing the model with more number of examples. Our study evaluate the suitability of GPT models with N-shot learning in a sensitive domain of healthcare sector leveraging wellness dimensions. We discover that a more domain-specific and contextually relevant training approach is necessary for effective classification.

We further consider fine-tuning all the four variants of GPT-3 [41] model: Ada, Davinci, Curie, and Babbage, to determine the effectiveness of Generative AI for domain-specific multi-label classification task in psychology-grounded healthcare sector. During the fine-tuning process, we train GPT-3 models for four epochs, using a batch size of 0.2% of the training data, and kept all other parameters at their default values. Once the models were fine-tuned, we compare their predictions to the ground-truth using a range of metrics.

B. 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. Recall indicates the percentage of relevant instances that are correctly identified for a particular wellness dimension. F1 score provides an overall measure of the model's effectiveness in correctly identifying relevant instances. Accuracy is the overall percentage of correctly classified instances across all wellness dimensions. For our task, F1-score is more important measures than accuracy is because the objective of this task is to correctly identify relevant instances for each wellness dimension, rather than to accurately classify all instances.

The Matthews Correlation Coefficient (MCC) measures the quality of classifications task. MCC produces a score between -1 and 1, where 1 represents a perfect prediction, 0 represents a random prediction, and -1 represents a total disagreement between prediction and observation. In MULTIWD, an imbalanced dataset, the number of instances belonging to each class varies significantly, and a classifier that predicts only the majority class can achieve high accuracy, but still perform poorly in identifying minority classes which can be verified using MCC score.

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. Area under the precision recall (PR) curve (AUPRC), on the other hand, measures the trade-off between precision and recall at various classification thresholds. It plots precision against recall, representing the overall quality of classifier's ranking of positive examples. In our multi-label classification task, where each instance may have multiple wellness dimensions, the accuracy alone may not be sufficient for evaluating a model's ability to correctly predict each label. Therefore, AUROC and AUPRC provide a more nuanced view of the model's performance across all labels.

TABLE V
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	43.96	80.89	55.70	01.22	26.66
One-shot learning	44.82	68.62	51.18	00.15	09.06
Few-shot learning	61.45	73.23	63.70	18.26	45.36
GPT-3 Ada	75.13	76.94	**75.94**	**41.55**	65.19
GPT-3 Baggage	74.43	76.36	75.36	40.18	64.41
GPT-3 Curie	74.61	76.44	75.44	38.66	64.73
GPT-3 Davinci	75.53	76.44	75.86	39.73	**65.39**

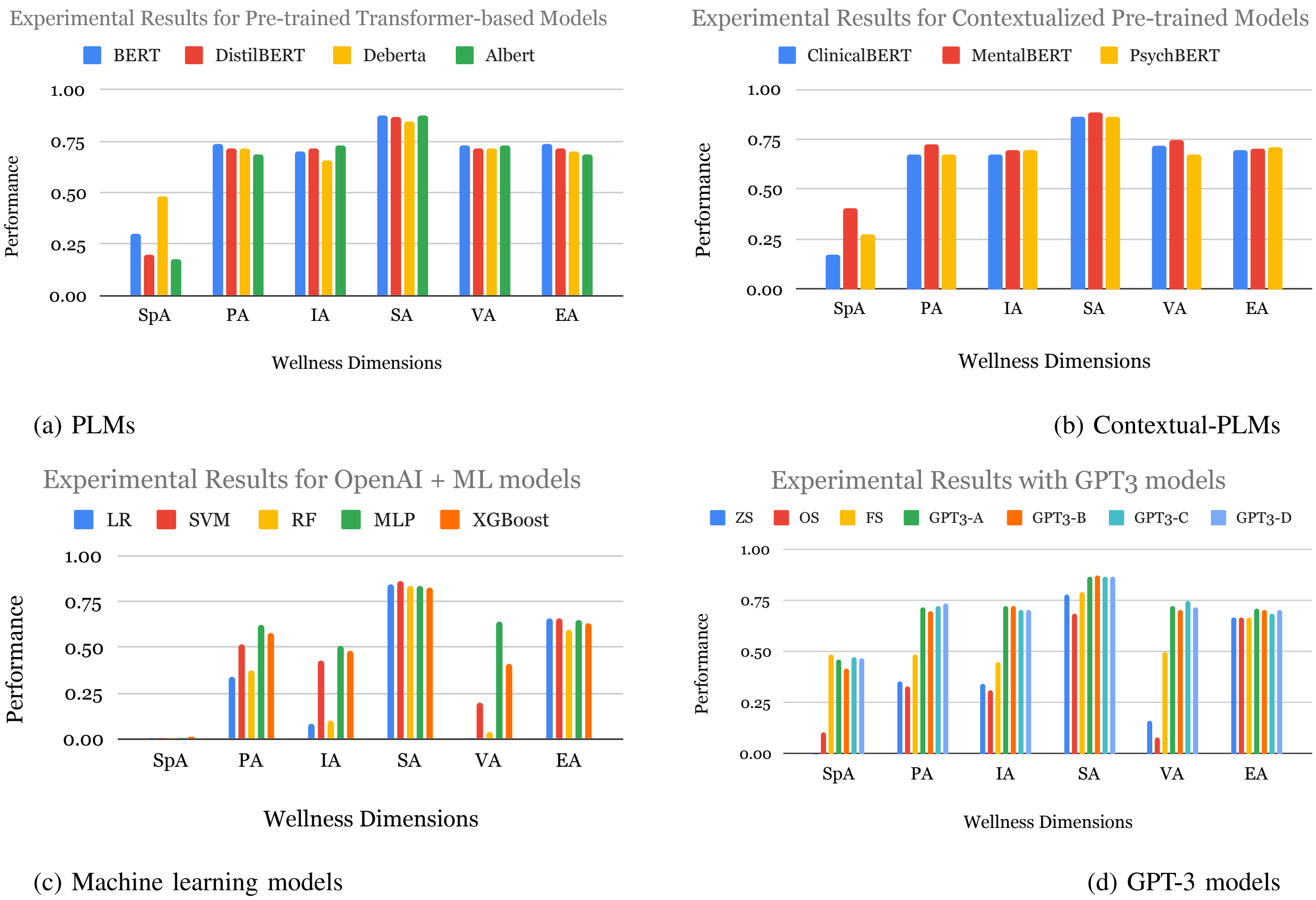
C. Experimental Results

We now discuss the experimental results that we present for **classifers as baselines** in the form of precision, recall, F1-score, Accuracy and MCC as illustrated in Table V. 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.

1) Classifiers as Baselines: 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 V. 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.

The GPT-3 Ada outperformed the other classifiers, achieving the highest F1-score of 75.94% and highest accuracy of 41.55% which is comparable to the accuracy obtained my BERT-base model. The success of GPT-3 Ada could be attributed to its adaptive learning approach, which allows the model to dynamically adjust its learning rate based on its performance on the training data. Its ability to perform efficiently and generalize to the new data demonstrates its superior performance compared to other GPT-3 variants. We find PLMs and GPT models as the most efficient models due to high MCC, accuracy and F1-score. Across all models, the accuracy scores were comparatively lower when compared to the F1 measure. 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.

2) Fine-grained Aspect based Analysis: We further illustrate the performance of all classifiers for each wellness dimension in Figure 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



(a) PLMs

(b) Contextual-PLMs

(c) Machine learning models

(d) GPT-3 models

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

offered by Open AI embeddings deployed to capture Wellness dimensions.

3) AUROC and AUPRC curves: The Figure 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.

D. Error Analysis

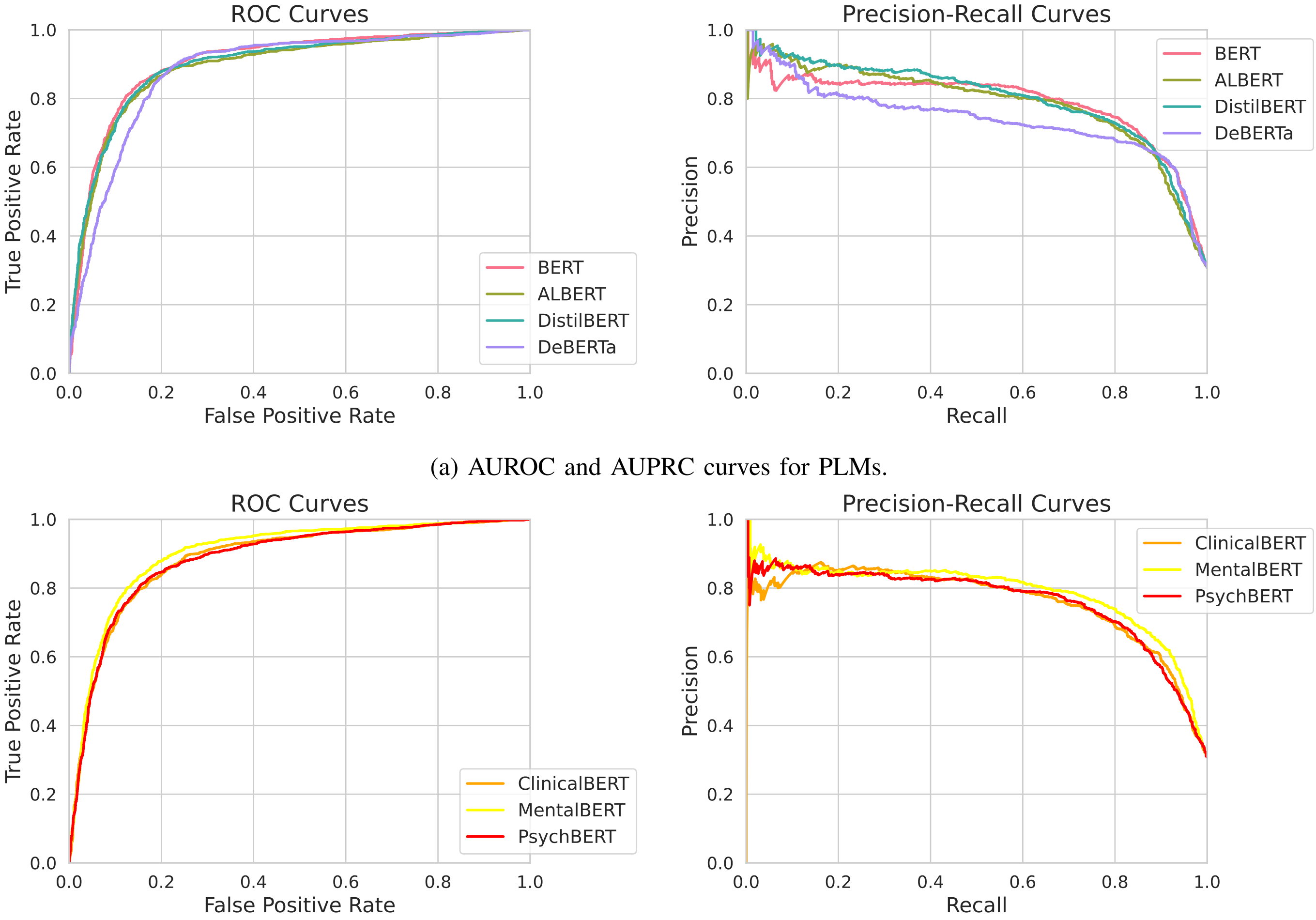
One of the key challenges is the need to incorporate domain-specific psychology for developing safe and responsible AI models. Thus, a deep understanding of the psychological concepts, clinical questionnaire and regular discussion with mental health practitioners are required to analyse the wellness dimensions in social media data. In addition to the need for domain-specific knowledge, there are a number of NLP-centered challenges that must be addressed such as:

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.

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.

The semantic ambiguity of words poses a challenge in decision-making for AI models because it can result in incor-



(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.

rect classification of posts into different wellness dimensions. To overcome this challenge, it is important to develop models that can accurately interpret the context in which words are used, and incorporate domain-specific knowledge to disambiguate words that have multiple interpretations. Additionally, incorporating human annotation and feedback in the model training process may improve the accuracy and effectiveness of the AI models.

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.

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. Overall, attention and ambiguity play a significant role in identifying wellness dimensions in social media data, and careful analysis is required to ensure accurate and effective categorization.

4) Informal Text: Social media users often use informal language and it is difficult to accurately process and analyze informal language in the context of wellness dimensions. The unstructured nature of social media text can make it difficult to accurately incorporate external knowledge, as the language used in social media posts may be informal, colloquial, or even contain slang or regional dialects. 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.

E. 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. 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. Furthermore, we have made our dataset and source code available on Github, enabling the baseline results to be easily reproduced.

VI. CONCLUSION

Our work is the first of its kind to conduct a quantitative analysis that emphasizes the necessity of identifying 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 instances in Reddit corpus. Through our experiments with traditional multi-label classifiers, we demonstrate the potential for improving 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. In future, we plan to infuse external knowledge to develop a more comprehensive multi-label classificaion approach. In order to facilitate future research and development, we plan to expand our dataset and explore data augmentation techniques to create a well-balanced dataset.

ACKNOWLEDGEMENT

We extend our sincere acknowledgement 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. Additionally, we extend our heartfelt appreciation to Prof. Sunghwan Sohn for his consistent guidance and support.

REFERENCES

[1] UN. Transforming our world: The agenda 2030 for sustainable development, 2015.

[2] UN ESCAP, World Health Organization (WHO, et al. Sdg 3 goodhealth and well-being: ensure healthy lives and promote well-being for all at all ages. 2021.

[3] Halbert L Dunn. High-level wellness for man and society. *American journal of public health and the nations health*, 49(6):786–792, 1959.

[4] Baocheng Pan, Hongyu Wu, and Xianhua Zhang. The effect of trait mindfulness on subjective well-being of kindergarten teachers: The sequential mediating roles of emotional intelligence and work–family balance. *Psychology Research and Behavior Management*, pages 2815–2830, 2022.

[5] Alison G Abraham, Chris Hong, Jennifer A Deal, Brianne M Bettcher, Victoria S Pelak, Alden Gross, Kening Jiang, Bonnielin Swenor, and Walter Wittich. Are cognitive researchers ignoring their senses? the problem of sensory deficit in cognitive aging research. *Journal of the American Geriatrics Society*, 2023.

[6] L Welty, A Harrison, K Abram, N Olson, D Aaby, and K McCoy. Substance abuse and mental health services administration.(2017). key substance use and mental health indicators in the united states: results from the 2016 national survey on drug use and health (hhs publication no. sma 17-5044, nsduh series h-52). rockville, md: Center for behavioral health statistics and quality. *Substance Abuse and Mental Health Services Administration. Retrieved. College of Health Sciences*, 106(5):128, 2019.

[7] Thomas E Joiner Jr, Kimberly A Van Orden, Tracy K Witte, and M David Rudd. *The interpersonal theory of suicide: Guidance for working with suicidal clients.* American Psychological Association, 2009.

[8] Adam Tsakalidis, Jenny Chim, Iman Munire Bilal, Ayah Zirikly, Dana Atzil-Slonim, Federico Nanni, Philip Resnik, Manas Gaur, Kaushik Roy, Becky Inkster, et al. Overview of the clpsych 2022 shared task: Capturing moments of change in longitudinal user posts. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 184–198, 2022.

[9] Tulika Saha, Vaibhav Gakhreja, Anindya Sundar Das, Souhitya Chakraborty, and Sriparna Saha. Towards motivational and empathetic response generation in online mental health support. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2650–2656, 2022.

[10] Ahmad Ishqi Jabir, Laura Martinengo, Xiaowen Lin, John Torous, Mythily Subramaniam, and Lorainne Tudor Car. Evaluating conversational agents for mental health: Scoping review of outcomes and outcome measurement instruments. *Journal of Medical Internet Research*, 25:e44548, 2023.

[11] Surjodeep Sarkar, Manas Gaur, L Chen, Muskan Garg, Biplav Srivastava, and Bhaktee Dongaonkar. Towards explainable and safe conversational agents for mental health: A survey. *arXiv preprint arXiv:2304.13191*, 2023.

[12] Muskan Garg, Chandni Saxena, Sriparna Saha, Veena Krishnan, Ruchi Joshi, and Vijay 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*, pages 6387–6396, 2022.

[13] Glen Coppersmith, Mark Dredze, and Craig Harman. Quantifying mental health signals in twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 51–60, 2014.

[14] Umashanthi Pavalanathan and Jacob Eisenstein. Emoticons vs. emojis on twitter: A causal inference approach. *arXiv preprint arXiv:1510.08480*, 2015.

[15] Ayah Zirikly and Mark Dredze. Explaining models of mental health via clinically grounded auxiliary tasks. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 30–39, 2022.

[16] Gunjan Ansari, Muskan Garg, and Chandni Saxena. Data augmentation for mental health classification on social media. *arXiv preprint arXiv:2112.10064*, 2021.

[17] Amir Hossein Yazdavar, Mohammad Saeid Mahdavinejad, Goonmeet Bajaj, William Romine, Amit Sheth, Amir Hassan Monadjemi, Krishnaprasad Thirunarayan, John M Meddar, Annie Myers, Jyotishman Pathak, et al. Multimodal mental health analysis in social media. *Plos one*, 15(4):e0226248, 2020.

[18] Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. Clpsych 2015 shared task: Depression and ptsd on twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39, 2015.

[19] Guangyao Shen, Jia Jia, Liqiang Nie, Fuli Feng, Cunjun Zhang, Tianrui Hu, Tat-Seng Chua, and Wenwu Zhu. Depression detection via harvesting social media: A multimodal dictionary learning solution. In *IJCAI*, pages 3838–3844, 2017.

[20] Andrew Yates, Arman Cohan, and Nazli Goharian. Depression and self-harm risk assessment in online forums. In *EMNLP*, 2017.

[21] Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. Smhd: a large-scale resource for exploring online language usage for multiple mental health conditions. In *27th International Conference on Computational Linguistics*, pages 1485–1497. ACL, 2018.

[22] David E Losada, Fabio Crestani, and Javier Parapar. Overview of erisk: early risk prediction on the internet. In *International conference of the cross-language evaluation forum for european languages*, pages 343–361. Springer, 2018.

[23] Lei Cao, Huijun Zhang, Ling Feng, Zihan Wei, Xin Wang, Ningyun Li, and Xiaohao He. Latent suicide risk detection on microblog via suicide-oriented word embeddings and layered attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1718–1728, 2019.

[24] Elsbeth Turcan and Kathleen McKeown. Dreaddit: A reddit dataset for stress analysis in social media. *arXiv preprint arXiv:1911.00133*, 2019.

[25] Han-Chin Shing, Philip Resnik, and Douglas W Oard. A prioritization model for suicidality risk assessment. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8124–8137, 2020.

[26] Ayaan Haque, Viraaj Reddi, and Tyler Giallanza. Deep learning for suicide and depression identification with unsupervised label correction. In *International Conference on Artificial Neural Networks*, pages 436–447. Springer, 2021.

[27] Soumitra Ghosh, Swarup Roy, Asif Ekbal, and Pushpak Bhattacharyya. Cares: Cause recognition for emotion in suicide notes. In *European Conference on Information Retrieval*, pages 128–136. Springer, 2022.

[28] Pravin Kumar and Manu Vardhan. Pwebsa: Twitter sentiment analysis by combining plutchik wheel of emotion and word embedding. *International Journal of Information Technology*, pages 1–9, 2022.

[29] Muskan Garg. Mental health analysis in social media posts: A survey. *Archives of Computational Methods in Engineering*, pages 1–24, 2023.

[30] Sungkyu Park, Inyeop Kim, Sang Won Lee, Jaehyun Yoo, Bumseok Jeong, and Meeyoung 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*, pages 557–570, 2015.

[31] Gabriel Roccabruna, Steve Azzolin, and Giuseppe Riccardi. Multi-source multi-domain sentiment analysis with bert-based models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 581–589, 2022.

[32] Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282, 2012.

[33] Adane Nega Tarekegn, Mario Giacobini, and Krzysztof Michalak. A review of methods for imbalanced multi-label classification. *Pattern Recognition*, 118:107965, 2021.

[34] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[35] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations, 2020.

[36] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020.

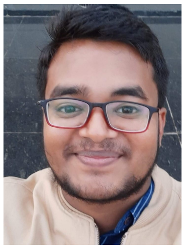
[37] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention, 2021.

[38] Vedant Vajre, Mitch Naylor, Uday Kamath, and Amarda Shehu. Psychbert: A mental health language model for social media mental health behavioral analysis. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1077–1082, 2021.

[39] Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. Mentalbert: Publicly available pretrained language models for mental healthcare, 2021.

[40] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission, 2020.

[41] OpenAI. Fine-tuning guide. https://platform.openai.com/docs/guides/fine-tuning, 2021. Accessed: May 11, 2023.



MSVPJ Sathvik (Student Member, IEEE) is an undergraduate student at Indian Institute of Information Technology Dharwad, Karnataka, India. His interests lies in application areas of *AI for social good* and *Large Language Models*. He has worked in mental health, fake news detection and other NLP-centered research projects.

Mr. Sathvik is a student member of Association for Computing Machinery (ACM) and Association for Computational Linguistics(ACL).

Dr. Muskan Garg , a distinguished postdoctoral research fellow at Mayo Clinic in Rochester, Minnesota, is a highly accomplished professional with a solid academic foundation and vast expertise in natural language processing (NLP), information retrieval, and social media analysis. She has completed her doctorate from Panjab University, India and has 10+ and 2+ years of research and teaching experience.

Dr. Garg's extensive publication record speaks to her scholarly achievements, with over 30 peer-reviewed papers published in prestigious journals such as Elsevier, Springer, and Emerald Insight. Her research findings have also been presented at renowned conferences including ACL, LREC, ICONIP, and NLDB, showcasing the recognition and impact of her work.