Highlights

WELLXPLAIN: Wellness Concept Extraction and Classification in Reddit Posts for Mental Health Analysis

Muskan Garg

- Introducing the need of datasets for reliable simulations in mental healthcare.
- Corpus construction for wellness concept extraction and classification.
- Analyzing domain-specific transformers and large language models for this task.
- Examining reliability of traditional multi-class classifiers.

WELLXPLAIN: Wellness Concept Extraction and Classification in Reddit Posts for Mental Health Analysis

Muskan Garg^{a,*,1}

^aMayo Clinic, Rochester, 55901 MN, USA

ARTICLE INFO

Keywords: Corpus construction mental health WellXplain wellness dimensions

ABSTRACT

Amid the ongoing mental health crisis, there is an increasing need to discern possible signs of mental disturbance manifested in social media text. Neglecting multi-dimensional aspects of social and mental well-being (i.e., wellness dimensions) over time can adversely affect an individual's mental health. During in-person therapy sessions, manual efforts are used to identify the causes and consequences of triggering latent factors of mental disturbance, which is a meticulous and time-consuming task for mental health professionals. To enable such fine-grained mental health screening, we define the task of determining wellness dimensions in Reddit posts as wellness concept extraction and classification problem. We construct a novel dataset called WELLXPLAIN, which consists of 3,092 instances and a total of 72,813 words. Our experts developed an annotation scheme and perplexity guidelines for annotation based on a well-adapted Halbert L. Dunn's theory of wellness dimensions. Further, the data encompasses human-annotated text spans as pertinent explanations for decision-making during wellness concept classification. We anticipate that releasing the dataset and evaluating the baselines will facilitate the development of new language models for concept extraction and classification in healthcare domain.

1. Introduction

A clinically significant impairment in a person's intellect, emotional control, or behavior is what is known as a mental disorder, suggesting cognitive decline. The UN Resolution of "Transforming our World: the Agenda 2030 for Sustainable Development" adopted in September 2015 [1], outlined an ambitious vision to tackle *Goal 3: Ensure healthy lives and promote well-being for all at all ages* of Sustainable Development Goals (SDG). By 2030, UN plans to reduce by one-third premature mortality from non-communicable diseases through prevention/treatment and promote the mental health and well-being. Untreated depression is conjectured to be the leading cause of suicide [2]. Reports released in August 2021 indicate that 1.6 million people in England were on waiting lists to seek professional help with mental health care.

Despite significant technological advances, mental health assessment remains a dark mark on public health efforts. Causes for mental health disturbances are broad, including an unspecific gamut of factors such as physical health problems, social conflicts (e.g., bullying, prejudice, stigma, race issues), abuse, grief, financial and professional difficulties, etc.. These causes are aggravated when patients do not disclose their concerns to mental health professionals, rather find solace on social media [3]. Motivated with non-intrusive high value information, social media data lessens the effect of limited availability of mental health practitioners. In these dire circumstances, online platforms are frequently relied

garg.muskan@mayo.edu (M. Garg)
ORCID(s):

upon not only as open and unobtrusive sources of information, but also as a place for honest disclosure, where people may freely express themselves along with their thoughts, beliefs, and emotions [4].

However, social media is extremely noisy because of popular culture references and slang terms that are pervasive in online expressions. This noise makes it hard to develop automated methods for mental health screening methods that levels with mental health professionals [5, 6]. Furthermore, prior work on mental health analyses from social media focuses on assessing posts that already exhibit particular mental health traits (e.g., analyzing mental health subreddits related to suicidality and depression) [7, 8, 9].

Amid the huge social impact of COVID-19 pandemic, the research community witness the presence of mental disorders in an individual through consequential affect on wellness dimensions due to prevailing reasons behind mental disturbance. We do not intend to invalidate the prior works with causal analysis such as CAMS, but expect to support the existing works with clinical concept extraction through our newly proposed wellness dimensions dataset WELLXPLAIN. We formulate this problem as a multivariate task to construct and release a dataset to develop a comprehensive and contextual AI models that determines all the wellness dimensions that are being affected, in a given Reddit post.

Social Media. We focus on Reddit, as it appeals to a significant number of subscribers due to its anonymity feature. For instance, r/depression subreddits has 934K and r/suicidewatch has 401K subscribers. These communities have enabled individuals to express their personal experiences and seek support for their mental conditions. The online users may suffer mental disturbances long before explicitly writing about it. Such individuals may be helped

¹https://www.un.org/development/desa/disabilities/envision2030-goal3 html

 $^{^2} https://www.theguardian.com/society/2021/aug/29/strain-on-mental-health-care-leaves-8m-people-without-help-say-nhs-leaders$

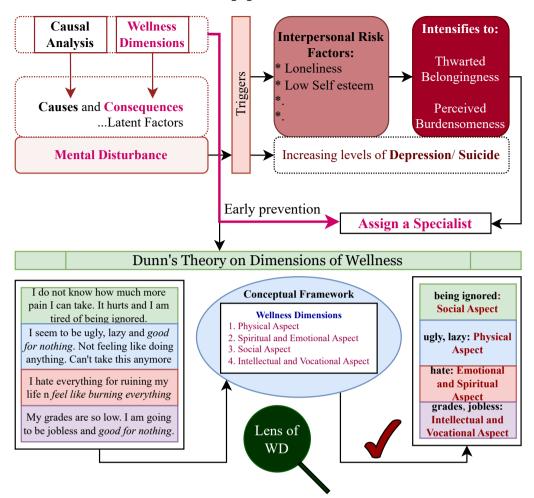


Figure 1: An illustration of the utility of Wellxplain. The intrinsic auxiliary tasks of wellness identification influence Al model's focus. We simulate the process of the early wellness concept extraction that undergo initial screening of a given Reddit post to determine the wellness concepts impacting mental health. The wellness dimensions in our work are grounded through Dunn's theory of wellness.

through early detection of negative mental health outcomes and subsequent intervention. This critical situation underscores the need for techniques to identify potential wellness concerns as clinical concepts in self-narrated texts submitted online, so as to be able to point out pertinent sources of professional counsel and therapy.

Motivated by the need for automation in mental health-care over social media resource points, we expand existing AI-driven models towards clinical concept extraction through wellness dimensions in human writings. Without such systems, a cry for help may remain unnoticed and ignored, or it may receive an overly generic message rather than one that addresses the specific mental health aspect. However, recent studies suggest that with the increasing deployment of AI-driven models on social media data, there is a growing risk of people turning to them as an outlet for disclosing personal misfortune [10, 11]. To this end, we further incorporate the clinical settings for clinical concepts extraction in self-narrated texts on Reddit.

Due to increased demand for quality mental healthcare and limited availability of mental health professionals, there exists a need to simulate the practice of identifying causes and consequences behind mental disturbance. Prior work has focused on coarse-granular mental health classification, which ignore a diverse array of underlying issues[12, 13]. According to the survey, individuals who lack strong familial and friendship ties are ten times more prone to mental health challenges³. The Interpersonal-Psychological Theory of Suicide [14] states that a serious suicidal behavior is rooted in the combination of acquired capability, Perceived Burdensomeness (PBU: hallucination of being disconnected/ burden on society), and Thwarted Belongingness (TBE: feeling of being isolated), which are collectively necessary and sufficient proximal causes evolving from affected Wellness Dimensions (see Figure 1). As such, we choose to simulate the problem of identifying clinical concepts in the form of wellness dimensions for Reddit posts reflecting suffering and distress. This task is well rooted in multiple clinicallygrounded applications including:

³https://mentalstateoftheworld.report/

- 1. Fine-grained Natural Language Understanding: A fine-grained NLU can discern wellness from user inputs, like identifying social aspect from not just explicit statements ("I am lonely") but also implicit ones ("I don't feel like getting out of bed but want to make friends"). It can understand the depth and subtleties in a user's language, such as differentiating between transient moods and persistent wellness states or detecting masked concepts behind seemingly deteriorating mental health. The model would also consider the context in which certain phrases are used to represent the wellness dimension, ensuring that it doesn't misinterpret casual uses of terms that might have psychological implications. With the capability to understand nuances, the system becomes more sensitive to users' psychological states, potentially identifying issues before they escalate. The deep understanding facilitates a more individualized therapeutic experience, accounting for the unique ways in which people express mindfulness.
- 2. Clinical Concept Extraction: Clinical concept extraction leverages NLP methods to autonomously detect and pull medically pertinent details from nonstructured textual data. Algorithms for extraction can be tailored to pinpoint and highlight particular keywords or expressions, especially those indicating the reasons (wellness dimensions) leading to outcomes (cognitive decline) like "lonely", "headache", or "poor grades". Automated platforms can alert human oversight teams or moderators about posts suggesting significant risk, ensuring faster response. By discerning dominant topics or challenges in such communities, moderators can compile resources, posts, or actions that cater to these specific needs.
- 3. Easy-to-reach-out 'counselors' through early detection of cognitive decline: Many individuals may not always recognize the need to reach out or might hesitate to approach professionals. Thus, having an automated system that identifies potential wellness issues in their social media posts can bridge this gap and offer timely assistance. The AI-driven model shall detect keywords, phrases, patterns, or sentiments that suggest a potential wellness dimension that needs attention. Rather than waiting for someone to acknowledge and vocalize their issues, the system identifies potential problems and suggests intervention. Addressing wellness concerns at an early stage can lead to better outcomes and prevent issues from escalating. The automated nature can provide a non-judgmental space for users to express their feelings. AI-driven models can learn and adapt based on individual user data, offering more personalized guidance over time.
- 4. Social Determinants of Health: Social Determinants of Health (SDOH) refer to the conditions in which people are born, grow, live, work, and age, including the wider set of forces and systems shaping the conditions of daily life. When applied to subreddits

such as r/depression and r/suicidewatch, the SDOH can help in understanding the underlying societal and environmental factors contributing to mental health and wellness. The SDOH provide a lens to understand the external factors that influence the multifaceted nature of wellness. While many users may post about their feelings or emotional states, digging deeper to understand the societal influences on these feelings can lead to more holistic support and interventions. For instance, Experiencing discrimination, social isolation, or a lack of social support can profoundly affect emotional well-being. Positive community ties, on the other hand, can enhance emotional resilience.

Contribution. We introduce an AI-based system for wellness dimensions that uses techniques to identify the concerns affecting wellness dimensions that may require counseling and therapy. The goal of such a system is to provide an efficient and effective way to identify individuals who may be struggling with wellness dimensions and to guide them towards appropriate support and resources. A recent surge in quantifying wellness dimensions suggest low-level analysis of personal writings for depressive symptoms in lieu of any of the six dimensions of wellness: different dimensions of wellness as: (i) physical aspect, (ii) intellectual and vocational aspect, (iii) social aspect, (iv) emotional and spiritual aspect. Such stressful situational aspects, prevailing for a long duration, may result in depressive symptoms. To this end, we consider the affected wellness dimensions (situational effects) as a major consequence of mental disturbance and propose the analysis of wellness dimensions grounded in established theory of High-level Wellness by Halbert L. Dunn [15], so as to investigate the evolution of mental health disturbances in social media posts. According to Dunn, high-level wellness is a state of complete physical, mental, and social well-being, beyond just the absence of disease or infirmity. He argued that high-level wellness is a dynamic process that requires ongoing effort and attention to maintain. Consider the following example:

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

Here, the author expresses three different situations, including problems with relationship status (Social Aspect), problems with work and career (Intellectual and Vocational Aspect), and loss of their beloved child (Social Aspect). However, the major focus of this text is the *loss of a baby boy*, emphasizing a major impact of life circumstances on the *social aspect*. We shall discuss such perplexities in more detail further below. These may pertain to a number of *dimensions of wellness* studied in psychology. For, to truly address and attenuate mental disturbances, we need datasets that can bridge the gap between life's challenges

Dataset	Details	Avail.
CLPsych [16]	Three types of annotated information using Depression, Control, and PTSD	S
MDDL [17]	300 million users and 10 billion tweets in Depression, Non-depression, and Depression candidate	Α
RSDD [18]	Reddit dataset of 9210 users in depression and 1,07,274 users in control group	ASA
SMHD [19]	Reddit dataset for multi-task mental health illness	ASA
eRISK [20]	Early risk detection by CLEF lab about problems of detecting depression, anorexia and self-harm	A
Sina Weibo [21]	3,652 (3,677) users with (without) suicide risk from Sina Weibo	AR
SRAR [13]	Posts from 500 Redditors (anonymized) & annotated by domain expert	ASA
Dreaddit [22]	190K Reddit posts of 5 different categories	A
GoEmotion [23]	Manually annotated 58k Reddit comments for 27 emotion categories	Α
UMD-RD ⁴ [24]	11,129 users who posted on r/SuicideWatch and 11,129 users who did not	ASA
SDCNL [25]	Reddit dataset of 1895 posts of depression and suicide	Α
CAMS [26]	Interpretable causal analysis of mental illness in social media (Reddit) posts	Α
WellXplain	Wellness concept extraction in Reddit posts	А

Table 1Different mental health datasets and their availability. A: Available, ASA: Available via Signed Agreement, AR: Available on Request for research work

and potential mental health disorders. This will expedite early detection and intervention. By understanding and identifying these wellness concerns as clinical concepts, we can pave the way for more accessible counseling resources, offering immediate support to those grappling with distress. As such, our task suggests the new research direction of extracting cause-and-consequence in a given text where cause (wellness concept extraction) aids in early detection of consequences (cognitive decline/ mental health illness).

While there are publicly available textual mental health assessment datasets [27], enabling AI models to account for wellness dimensions may yield rich and fine-grained contextual information, ultimately also relevant to clinicians (see Table 1). To this end, we observe a scarcity of available datasets for the finer-grained task of identifying clinical concepts through wellness dimensions as a consequence of poor well-being. To the best of our knowledge, existing literature has no language resources on wellness dimensions that are available for research and development. The reason behind limited public availability is due to (i) limited lowlevel analysis for cause-and-consequence of mental health and (ii) sensitive nature of the data. The public release of our dataset on GitHub⁵ shall facilitate future research by raising the red alert on determining one of the pre-defined concepts (wellness dimensions) affecting mental disturbance via a longitudinal study of Reddit posts [28]. Thus, our key contributions can be summarized as follows:

- To the best of our knowledge, this is the pioneer quantitative investigation emphasizing the importance of concept extraction using wellness dimensions to analyze mental health in Reddit posts.
- 2. We construct and release WELLXPLAIN, a new English dataset of 3,092 instances with 72,813 words as a *multi-class classification* problem with *text-span* explanations.

 Our experiments with machine learning classifiers, sequence-to-sequence models, Transformer models and large language models set the baseline and suggest the opportunities for further research and development.

2. Corpus Construction

The current landscape offers abundant open-source datasets for research and application development across various fields. However, the scarcity of high-quality datasets is a pressing issue, particularly in the rapidly evolving field of Natural Language Processing (NLP), where they are vital for training and evaluating Language Models.

Real-world datasets are often complex, unorganized, and lack structure, and the performance of models is closely tied to the dataset's quantity, quality, and relevance. It's crucial to understand a dataset's nature, significance, and its foundational role in building artificial intelligence-based NLP systems. Bearing this in mind, the principal obstacle in dataset creation lies in safeguarding the uniformity and trustworthiness of the annotations. To surmount this challenge, we construct the psychology-driven annotation scheme using Dunn's theory of wellness dimensions and experts' intervention for annotating dataset and deploy it. In this section we first discuss the data acquisition, followed by annotation scheme design and annotation task.

2.1. Data Acquisition

Reddit is a social media platform for open discussions and its ability for individuals to post anonymously make it a diverse platform for candid and personal data collection about mental health complexities based on personal experiences. Being anonymous, people express their thoughts and share their experiences about mental disturbance with ease, often facilitated by a non-judgmental environment as compared to other prominent social media platforms. Although the Python Reddit API Wrapper (PRAW) API,

 $^{^{5} \}verb|https://github.com/drmuskangarg/WellnessDimensions/|$

an interface used for data collection from Reddit, allows the retrieval of information about authors and creation date, we limit the information to the title and text of the post to maintain the privacy of data. The Reddit platform provides an opportunity of sharing user-generated and usercurated context in the community-specific subreddits for open discussions indicating special themes or conditions. In the past five years, the subreddits associated with mental disorders such as r/depression and r/SuicideWatch have grown more than 100% and 250%, respectively. A limit on the number of characters is 10,000 for a comment, 40,000 for a post, and 300 for post titles. Prior work on mental health analysis in Reddit posts suggests the use of next word prediction models, which are often computationally expensive for lengthy posts. Although we suggest the need of a more comprehensive and contextual approach to solve the problem, we keep the maximum length of the Reddit posts at 256 tokens in our dataset to facilitate machine learning research. The use of NLP for Reddit posts has major challenges of double negation, syntactic and semantic ambiguity, as well as commonsense and domain specific knowledge infusion.

To limit the length of a Reddit post, we limit the maximum number of words in each post to 256. We performed expert-guided manual cleaning and filtered the Reddit posts to limit them to *user experience* impacting mental disturbance, resulting in 3,092 posts. Our three experts: a senior clinical psychologist, a rehabilitation counselor, and a social NLP researcher, frame annotation schemes and perplexity guidelines to annotate a given Reddit post with one of the pre-defined dimensions of wellness.

2.2. Annotation Scheme Design

Wellness Dimensions. Recent literature suggests six different dimensions of wellness as a consequence of mental and social well-being, namely, (i) physical aspect, (ii) intellectual aspect, (iii) vocational aspect, (iv) social aspect, (v) emotional aspect, and (vi) spiritual aspect [29, 30]. Our experts performed a pilot study on 40 instances, observing a high overlap and ambiguity of words and interpretations, respectively, in two different cases, (i) intellectual vs. vocational aspect, (ii) emotional vs. social aspect. The observations are supported by Dunn [15] in his model of *Wellness as a Conceptual Framework*, which defines wellness as an integrated balance of social, physical, cognitive (intellectual and vocational), and spiritual (emotional) health.

We further investigate the corresponding consequences of these cause on similar lines through the psychologically-driven theory of *dimensions of wellness* by Dunn, and define them as:

Physical Aspect (PA): Physical development encourages a good diet and nutrition, while discouraging the use of tobacco, drugs, and excessive alcohol consumption. Optimal wellness is met through good exercise,

good sleep, energetic behaviour, enthusiasm, and eating habits. Body shaming affects the physical well-being of a person, making them realize their problems with medical history and physical appearance.

- Intellectual and Vocational Aspect (IVA): The use of intellectual and cultural activities within or beyond the classrooms, combined with the human and learning resources. Wellness of a person cherishes intellectual growth and stimulation. The occupational dimension recognizes personal satisfaction and enrichment in one's life through work. It affects their attitude towards creative thinking, professional growth and other financial expenses.
- Social Aspect (SA): The social dimension emphasizes the interdependence between society and nature. A person becomes more aware of their importance in society as well as the impact which they may have on multiple environments. Social connections help in flourishing interpersonal factors of their nature and enhance the ability to emphasise cultural impacts.
- Spiritual and Emotional Aspect (SEA): The spiritual dimension recognizes the meaning and purpose in human existence. It includes the development of a deep appreciation for the depth and expanse of life and natural forces that exist in the universe. It is characterized by a peaceful harmony between internal personal feelings and emotions. The emotional dimension enhance the awareness and acceptance of one's feelings and the extent of positivism and enthusiasm about one's self and life. It includes the capacity to manage one's feelings and related behaviors including the realistic assessment of one's limitations, development of autonomy, and ability to cope effectively with stress.

This list is not exhaustive, but is a starting point for our study, giving rise to a final set of four dimensions, namely, (i) Physical Aspect (PA), (ii) Intellectual and Vocational Aspect (IVA), (iii) Social Aspect (SA), (iv) Spiritual and Emotional Aspect (SEA). Samples from the dataset are given in Table 2.

Guideline development. Our experts developed annotation guidelines by using the definitions of Wellness Dimensions given by Dunn [15] as mentioned above. Given that wellness dimensions are a highly subjective and complex issue, identifying them accurately can be challenging and prone to errors with naive judgment. To overcome this, our team of experts negotiates a trade-off between *using text-based marking* for developing advanced AI models and reading between the lines to provide psychological insights, while framing annotation schemes. Our experts developed annotation guidelines to aim at:

1. Finding potential text spans to identify dimensions of wellness affecting mental health.

Table 2
A sample table to present dimensions of wellness and annotated samples. Here, column 1 presents four WD: Wellness Dimensions, namely, PA: Physical Aspect, IVA: Intellectual and Vocational Aspect, SA: Social Aspect, and SEA: Spiritual and Emotional Aspect.

WD	Annotated samples	Text Span			
	My stomach bulges out too much, my face is fat, my acne is ugly, I need to shave, my teeth are yellow, my butts too small, etc	face is fat, acne is ugly, teeth are yellow			
PA	Well it did, to the point where any pressure on my stomach caused me pain and I would have to just lay down until the upset stomach subsided	upset stomach			
IVA	I've almost failed 3 classes the past 3 school years, and I'm on the brink of failing another one	failed 3 classes the past 3 school years			
	I'm a 23 year old unemployed woman still living with my mom	unemployed woman			
	Bad family, just got out of a bad relationship, have always been bad at making friends	bad at making friends			
SA	I'm 21, both parents dead, no family support, just friends, my brother abandoned me and left me out in the cold and kicked me out and cut me off because I was out drinking as a teenager	both parents dead, no family support, just friends, my brother abandoned me			
SEA	Failing half my classes, same as last semester, going to drop out of college at the end of the semester with no life direction, no ambition, no motivation, no desires, dreams, will to live	no ambition, no motiva- tion, no desires, dreams, will to live			
	I want someone to hold me and prove me wrong and tell me that I am valued and not worthless	worthless			

2. Developing a language resource to build explainable AI models for (content-aware domain-specific) 4-class classification problem of determining dimensions of wellness.

The experts annotated 40 samples of the dataset in isolation, using fine-grained guidelines to avoid biases [31]. Annotations are crucial in supervised machine learning, where you train a model based on labeled examples. 'Finegrained' suggests that these rules are very specific, aiming to capture nuances in the data. Halbert L. Dunn's concept of "High-Level Wellness" in 1959 is among the foundational works on the modern understanding of wellness. Dunn proposed that wellness is not merely the absence of disease but a proactive, holistic approach to life that combines the physical, mental, and social aspects, among others. By using Dunn's model of Wellness as a foundation for creating annotation rules, the project anchors its methodology in a wellestablished, researched framework. This can lend credibility to the endeavor, as it's not based on arbitrary or ad hoc criteria but on a historical and scholarly model. Using finegrained rules means that the project is making efforts to capture the subtleties and nuances of wellness as described by Dunn. For instance, instead of broadly labeling a piece of data as 'wellness-related', there might be more specific labels like 'emotional wellness', 'physical wellness', 'social wellness', etc., all derived from Dunn's conceptualizations.

One of the biggest challenges in AI and machine learning is ensuring that models do not inherit biases present in data or human annotators. By relying on a structured, well-defined framework for annotations, the aim is to provide a standardized basis for labeling data, minimizing subjective biases that could arise from individual interpretations. It's essentially a way to systematize and standardize the process, ensuring that different annotators would (ideally) label the same piece of data in the same way. A machine learning model's predictions are as good as the data it's trained on. If the training data is annotated with consistency, precision, and alignment with a recognized framework like Dunn's, the resulting model is more likely to make accurate and reliable predictions in line with that framework.

Therefore, using Dunn's model-based fine-grained rules for annotations helps ensure that the annotations are more objective and consistent, leading to more reliable and trustworthy results. The current annotation scheme has limitations in capturing all the aspects of the phenomenon. Despite these efforts, due to the subjective nature of this complex task, we encountered possible dilemmas. Therefore, we propose a set of perplexity guidelines that are intended to simplify the task and make future annotations easier to perform.

2.3. Perplexity guidelines

The task of annotating text for wellness dimensions can be complex, especially when the text mentions multiple reasons or contains ambiguity in its interpretation. In order to simplify the task and facilitate future annotations, we have developed a set of perplexity guidelines. These provide a framework for annotators to follow, ensuring that their annotations are consistent and accurate, thereby reducing the potential for errors or misunderstandings when interpreting the text. By using these guidelines, we made the task of annotating wellness dimensions more manageable and accessible, ultimately leading to a better understanding of the relationship between text and well-being. We enumerate the major perplexity guidelines as follows:

1. **Presence of Multiple Aspects**: Our team of experts has encountered posts on social media platforms where individuals express their feelings due to multiple reasons. To overcome this challenge, our team suggests a solution where the specific text spans in the post that contribute to the more focused health consequences should be identified and associated with the corresponding wellness dimension. This approach allows for a more nuanced annotation of the wellness dimensions and provides a clearer understanding of the underlying causes of mental health issues. Consider the following post *P*₁:

 P_1 : I cannot do anything without screwing it up, I just got suspended from school my family, my friends, all lose their trust in me, I'm just not cut out for anything I don't think the world has a place for me.

In the given post P_1 , suspension from school affects the IVA of the author, resulting in disturbed SA. This situational effect on wellness dimensions results in mental disturbance. For instance, in a given post P_1 , the affected dimension of wellness is SA.

2. **Annotation Ambiguity**: As wellness dimensions are not atomic in nature, identifying them is a highly subjective and complex task, which can be challenging for naive human annotators. Even experts may provide varying annotations for the same post, adding to the difficulty of the task. For instance, consider a post *P*₂:

 P_2 : I hate my home I hate my family and I hate my life I normally can ignore it but sometimes it just gets too much and I end up like I am now crying at my desk with no real reason why

Although there are multiple aspects in this post P_2 , we must consider the holistic aspect as per the experts' opinion. For instance, the above post is affecting SEA, as the user mentions their feelings about lifestyle.

3. Reading between the lines: The text may contain implicit or subtle hints that suggest a particular wellness dimension, which can be difficult for annotators to identify. We follow a strict guideline for text-span extraction-based classification and avoid making assumptions about the narrator. However, our team of experts has agreed that clear and meaningful words

that suggest one of the four wellness dimensions should be annotated accordingly. This approach ensures that the annotations are accurate and consistent, even in cases where the text may not explicitly mention a particular wellness dimension. For instance, consider the following post P_3 :

 P_3 : I think being bored 9-5 is even more depressing! Now I'm desperately struggling to keep happy and I have this low to medium level of anxiety that just won't go away no matter how hard I try and think differently.

The given post P_3 clearly mentions 9-5 which, as per commonsense, is interpreted as referring to working hours without the need for further assumptions. Thus, the monotonous work environment leads to an annotation of this post as IVA.

2.4. Annotation task.

To ensure consistency among annotators, the annotation and perplexity guidelines are used for providing formal training to all three student annotators. Our team trains three postgraduate students for manual annotations. Through three successive sessions of trial and error analysis, which involve the annotation of 40 samples per session, we aimed to achieve coherence among the annotators. After the training sessions, we asked the students to perform a two-fold annotation of each data point: (i) Assigning a Wellness label, and (ii) Identifying the text spans that explain the reason for the given label.

Inter-annotator agreement. In our endeavor to recognize wellness dimensions from textual data, we have delineated two specific annotation tasks:

- 1. **4-Class Classification Task**: The primary objective here is to classify a provided text into one of the four predetermined wellness dimensions. Given the nature of our dataset, which consists of relatively short data points, our initial approach involves multi-class classification. This not only aids in pinpointing the most influenced wellness dimension by the given text but also facilitates an assessment of the robustness and statistical properties of our dataset.
- Text-Spans Identification: This task zeroes in on identifying specific segments or spans of text that serve as indicative markers for the wellness dimensions. These spans provide a granular understanding, shedding light on which particular sections of the text resonate with the wellness dimension of interest.

To ensure the *reliability* of our annotations, particularly for the 4-class classification, we employed multiple validation measures. Initially, three independent student annotators were engaged in this task. Their annotations underwent a statistical evaluation using Fleiss' Kappa inter-observer

Table 3Statistics of WellXplain

Criteria	Frequency
Statistics	
Number of Posts	3092
Total number of Words	72,813
Max. number of Words	231
Total number of Sentence	3,376
Max. number of Sentences/post	7
Wellness Dimension	
Physical Aspect	750
Intellectual and Vocational Aspect	592
Social Aspect	1,139
Spiritual and Emotional Aspect	621

agreement, yielding a value of $\kappa = 74.39\%$. This value underscores a significant agreement among them, emphasizing the trustworthiness of their annotations.

To augment *accuracy*, we adopted a majority voting mechanism: for any given data point, if a label was endorsed by at least two out of the three annotators, it was considered as the final label. Yet, we introduced another layer of scrutiny. Recognizing the nuances and intricacies of the domain, these annotations were further vetted by a seasoned clinical psychologist. This ensured that our labels resonated with domain-specific expertise.

For a comprehensive perspective on the data distribution across the distinct wellness dimensions, we compiled a table detailing the statistics. This table, designated as WELLX-PLAIN, offers insights into the count of data points under each dimension and other relevant details (See Table 3).

For the task of text-span extraction, our three annotators, combined as a group, were asked to choose a set of words as text-spans representing explanations for corresponding wellness dimensions. The student annotators jointly extracted text spans as a potential *explanation* for each data point to form the final list of annotations for explanation. Our three experts carried out an agreement study over the final lists of explanations generated by a group of three student annotators using *Fleiss' Kappa* statistics [32], resulting in a κ score of 87.32%. The experts were asked to categorize their annotations as either Agree or Disagree. Although the inter-annotator agreement for the former task is slightly lower due to confusion between PA and SEA, we observe a higher agreement for the selection of explanatory text-spans.

2.5. Additional Data Analysis

We observe the number of samples for different aspects as: PA (24.26%), IVA (19.15%), SA (36.84%), and SEA (20.09%). The large number of samples for sentiment analysis (SA) indicates that our society may face more challenges in dealing with issues related to "near and dear ones" and "loneliness." These issues are often complex and difficult to handle, and may require more attention and resources to address effectively. The high number of samples in SA suggests that individuals are seeking emotional support and

 Table 4

 Frequent words in text spans for each wellness dimension.

WD	Avg.(W)Most Frequent Words						
PA	4.26	ugly, anxiety, sleep, meds, pain, tired, panic, drunk, alcohol, diagnosed					
IVA	5.46	job, school, work, college, money, failing, failed, time, life, year					
SA	4.86	friends, alone, family, lonely, people, feel, want, someone, parents, friend					
SEA	3.46	hate, feel, sad, worthless, motivation, anxiety, life, cry, shit, useless					

validation through social media platforms, which highlights the importance of addressing mental health concerns in these online spaces. The lower number of samples for IVA suggests that there is strong administrative control and good governance in this domain. This could be attributed to stringent regulations and protocols, or the expertise and experience of the individuals involved. The identification of IVA incidents may require specialized knowledge or expertise, which could explain the lower number of samples. Furthermore, the importance of interpersonal relationships outweighs that of intellectual and vocational skills. It is crucial to maintain healthy relationships with others, as they play a significant role in our overall well-being. While vocational and intellectual skills are important, they do not have the same impact on our mental health as our interactions with others. Developing and maintaining positive relationships can lead to a more fulfilling life and better mental health outcomes.

2.5.1. Frequent words in text spans.

The *text spans* marked or the *text-segments* in the original texts serve as explanations for decisions regarding wellness dimension selections. The explanatory text-spans are analyzed to identify the most frequent words for each aspect, and the results are tabulated in Table 4.

The data consists of segments of text termed as 'text-spans', which likely provide explanations or context about a topic. Analyzing these text-spans is a methodological approach to understand word patterns, occurrences, and the contexts in which words appear.

- 1. Tabulated Results: The results of this analysis, particularly the most frequently occurring words for each aspect (See Table 4) offers a breakdown of which words are the most predominantly associated with each wellness concept.
- 2. Unique Linguistic Signatures: Upon careful examination of the table, a noteworthy observation is that the lists of frequent words for each aspect, such as SA and SEA, are distinctly separate. This delineation suggests that unique linguistic signatures or patterns are associated with different wellness concepts.
- 3. Common Words Across Aspects: Interestingly, despite the uniqueness in word lists for each aspect,

- certain commonalities exist. For instance, the word "feel" finds its presence in multiple aspects. This revelation emphasizes that some words, despite their universality, are not exclusive indicators of a specific aspect.
- 4. Significance of Context: As exemplars, consider the phrases "feeling lonely" and "feeling useless". The former is categorized under the SA aspect, whereas the latter falls under SEA. Such examples accentuate the pivotal role context plays in language. The meaning and the ultimate categorization of a word can metamorphose based on its adjacent words or the overarching sentiment of the sentence.

The multifaceted nature of language emerges not just from individual words but their composite structures and contextual usage. The dynamic interplay of words in varied contexts poses inherent challenges for AI models, especially those striving for linguistic understanding or classification. Simple frequency-based paradigms might be inadequate, making a strong case for advanced contextual models. An undue focus on keyword frequency sans context could inadvertently lead to misclassification. This accentuates the need for nuanced analysis techniques.

3. Our Framework

3.1. Problem definitions

The Concept Extraction Framework (CEF) is meticulously designed to automate the process of extracting salient wellness concepts from expansive textual repositories. It seeks to not only localize specific concepts but also discern the intricate relationships or contextual embeddings in which these concepts reside. The mathematical characterization of this task is the multi-class classification challenge for categorizing a given text into one of the predefined wellness concepts.

3.2. Mathematical Notations and definitions

The multi-class classification of text into predefined wellness dimensions has emerged as a pivotal task in the realm of textual analysis. In this section, we introduce a mathematical framework for such multi-class classification. We define a classification function that maps a given text to its corresponding wellness dimension. This work leverages probability distributions to capture the likelihood of a text belonging to a particular wellness dimension. To fine-tune our model, we employ the categorical cross-entropy loss, aiming to minimize it across training instances. Optimization techniques, especially stochastic gradient descent, play a crucial role in updating our model parameters and achieving optimal performance.

3.2.1. 4-class classification task

Consider the following mathematical notations for the task of multi-class classification:

T: A given text or document that needs to be categorized.

- D: D is the complete set or list of all possible wellness dimensions that you want to classify the Reddit post into. Thus, the set of predefined wellness dimensions, where $D = \{d_1, d_2, \dots, d_k\}$ and k is the total number of wellness dimensions. Here, k = 4 for four different wellness concepts.
- *f*: A classification function which maps a given text *T* to one of the dimensions in *D*.

Objective: When a user posts something on Reddit (or a specific Reddit community focused on wellness), the goal of the classification model would be to automatically determine which dimension of wellness the post is discussing or relating to. Thus, given a Reddit post T, our multi-class classification task aims to determine the most appropriate wellness dimension d_i from D to which the text belongs.

$$f: T \to D$$
 (1)

$$f(T) = d_i (2)$$

where d_i is the wellness dimension that best represents the text T.

Probability Distribution: For many modern classification models, especially those based on deep learning, the output is often a probability distribution over the set of wellness concepts. Let P be the probability distribution vector generated by the model for text T over the set D, where:

$$P = \{p_1, p_2, \dots, p_k\} \tag{3}$$

Here, p_i represents the probability that the text T belongs to the wellness dimension d_i .

The final classification decision can then be made by selecting the dimension with the highest probability:

$$d^* = \arg\max_{d_i \in D} p_i \tag{4}$$

Where d^* is the most probable wellness dimension for text T

Loss Function: In order to train the model, a suitable loss function is required. For multi-class classification problems, the commonly used loss function is the categorical cross-entropy loss:

$$L(T, d_i) = -\sum_{i=1}^{k} y_i \log(p_i)$$
 (5)

Where:

- y_i is a binary indicator (0 or 1) if dimension d_i is the correct classification for T.
- p_i is the predicted probability that T belongs to dimension d_i .

Training Objective: The main objective during training is to minimize the loss function across all training examples. The parameters of the classification function f are updated iteratively using optimization algorithms, like stochastic gradient descent, to achieve this goal.

3.2.2. Text-span Identification

Attention mechanisms in deep learning allow models to focus on specific parts of the input data. For textual data, this essentially means focusing on certain text-spans or words. The attention weight associated with each word or text-span indicates its importance.

Let $T = \{w_1, w_2, \dots, w_n\}$ be the sequence of words in text T, where w_i represents the i^{th} word.

The attention mechanism assigns a weight a_i to each word w_i in T. The sequence of attention weights is given by $A = \{a_1, a_2, \dots, a_n\}$ where:

$$a_i = Attention(w_i)$$
 (6)

Identifying significant text-spans: In the pursuit of classifying texts into one of the four wellness dimensions, the importance of specific words or tokens in the text cannot be overstated. These words, which significantly influence the decision-making process of our model, form a unique subset of the textual data. The words I with the highest attention weights that enables decision-making for the task of 4-class classification is defined as:

$$I = \{ w_i \mid a_i > \theta, \forall i \} \tag{7}$$

where θ is a predefined threshold determined empirically in language models.

Comparison with Ground Truth: Let $G = \{g_1, g_2, \dots, g_m\}$ be the set of text-spans or words that represent the ground truth for the text-spans determining wellness concept in a given text T. The reliability R of the proposed model can be determined by comparing the set I with G:

$$R = \frac{|I \cap G|}{|G|} \tag{8}$$

Here, $|I \cap G|$ represents the number of words that are both identified as important by the attention mechanism and present in the ground truth. |G| is the total number of words in the ground truth. The reliability R gives a ratio of correctly identified important words to the total important words in the ground truth.

3.3. Large language models

GPT-3 is a state-of-the-art deep learning model designed for natural language processing tasks. It is the fourth iteration of the GPT architecture and has been trained on vast amounts of text data, making it highly capable of understanding and generating human-like text. The four model variants differ primarily in size and computational demand: (i) *Ada*: The smallest variant, suitable for tasks with limited computational resources, (ii) *Babbage*: Medium-sized, balancing computational demand and performance, (iii) *Curie*: Larger than Babbage, ideal for complex tasks needing higher accuracy, (iv) *Davinci*: The most powerful variant, used for highly demanding tasks but requires the most computational

resources. All variants process a given Reddit post to predict a probability distribution over the wellness dimensions. The core architecture comprises multiple transformer layers, with the output being classified into the desired wellness dimensions. Let:

- 1. T denote a given Reddit post.
- 2. *D* be the set of pre-defined wellness dimensions, where $D = \{d_1, d_2, d_3, d_4\}$.
- 3. M represent the set of GPT-3 variants used for classification, given by $M = \{Ada, Babbage, Curie, Davinci\}.$

For a given T, a model variant $m \in M$ predicts a probability distribution P_m over the set D. Here, $P_m = \{p_1, p_2, p_3, p_4\}$, where p_i indicates the probability that the text T belongs to the wellness dimension d_i . The classification decision for a model variant m is then:

$$d_m^* = \arg\max_{d_i \in D} p_i \tag{9}$$

Where d_m^* is the most probable wellness dimension for text T as predicted by model variant m.

By empirically testing all four variants, the objective is to gauge their performance in classifying Reddit posts into the pre-defined wellness dimensions. This would provide insights into which variant offers the best balance between computational cost and classification accuracy for this specific task.

3.4. Working Instance

One of the remarkable features of Transformer-based models like GPT-3 is their attention mechanism, specifically the self-attention mechanism. This allows the model to focus on different parts (text-spans) of an input sentence when producing an output. In the context of our classification task, this mechanism enables the model to weigh the importance of different text-spans in a Reddit post when deciding on the most appropriate wellness dimension.

1. Input:

A sample Reddit post, e.g., "Lately, I've been feeling a bit isolated from everyone. The loneliness is truly setting in."

2. Processing with Attention Mechanism:

- The model interprets the post using multiple attention heads. Each head concentrates on different text-spans.
- One attention head might emphasize the terms "feeling" and "loneliness", identifying them as signs of an social state.
- Another might give attention to the segment "isolated from everyone", providing emotional context in the post.

3. Output Probability Distribution:

The collective understanding from all attention heads results in a probability distribution over the wellness dimensions. The predicted distribution might be: PA: 0.08 IVA: 0.05 SA: 0.75 SEA: 0.12

The model, in this case, asserts with a high likelihood that the post is indicative of an 'Social Aspect'.

4. Visualization of Attention:

Advanced tools can visualize attention scores, accentuating the parts of the post that majorly influenced the model's conclusion. Words such as "feeling", "loneliness", and "isolated" might be intensified, illustrating their central role in the 'Social Aspect' dimension classification.

Understanding where the model is focusing allows insights into its decision-making process. This knowledge not only reveals model operations but can also aid users or developers to trust the model's results or pinpoint training needs.

4. Experiments and Evaluation

Extensive experiments for multi-class classification and explanation extraction were conducted to assess the effectiveness of the baselines and to identify their limitations.

4.1. Baselines

In the multifaceted landscape of multi-class classification, a plethora of algorithms and methodologies have been devised and employed. Our study embarks on a detailed examination of four distinct types of multi-class classifiers, spanning both classic machine learning and contemporary deep learning models.

4.1.1. Classic Machine Learning Algorithms

The very foundation of our investigation stems from classic machine learning algorithms. Here, the focus is on employing traditional, yet powerful, classifiers. Our text data is initially transformed into a numerical representation through the Term Frequency-Inverse Document Frequency (TF-IDF) technique. This representation encapsulates the importance of words in relation to the entire corpus. Logistic Regression (LR) is a statistical model used for predicting the probability of an instance belonging to a particular category. Given its robustness and simplicity, LR is often a go-to for many text classification tasks. As an ensemble learning method, Random Forest (RF) constructs a multitude of decision trees during training and outputs the class that is the mode of the classes of the individual trees. Owing to its ability to mitigate overfitting and handle large datasets with higher dimensionality, it's a prominent choice for various classification challenges.

4.1.2. Recurrent Models

Delving into the realm of deep learning, we focus on **recurrent models**. Given the sequential nature of textual data, Recurrent Neural Networks (RNNs) are apt choices as they are innately suited to handle sequences. To represent

the text, we employ pretrained word2vec embeddings. These embeddings capture semantic information and relationships between words. Within the recurrent paradigm, we deploy Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU). LSTM units are a type of RNN architecture. They possess the ability to remember patterns over long durations, thereby making them effective for sequence-based tasks. As per [33], GRUs are a variant of LSTMs. They come with a simpler structure, requiring fewer parameters and often achieving comparable performance. Both LSTM and GRU models in our study are constructed with two layers, comprising 32 hidden neurons each. The non-linear ReLU activation function is chosen to introduce non-linearity into the model. For the training process, we harness the power of the cross-entropy loss function, which serves as a good measure for classification tasks. Additionally, the Adam optimization algorithm, with a learning rate set at 0.001, is used to refine our model parameters and guide the training process to convergence.

4.1.3. Encoder only Transformers

Our investigation encompasses several variants of **encoder-only Transformers**. These models primarily focus on the encoder component of the original Transformer architecture. They are adept at capturing the context of every token in a sequence, making them particularly useful for tasks such as text classification, named entity recognition, and more. For our study, the input text is tokenized using a pretrained Transformer tokenizer, resulting in 768-dimensional vectors. These vectors serve as the initial embeddings for our models. The models we consider are:

- 1. **BERT** (**Bidirectional Encoder Representations from Transformers**): Introduced by [34], BERT is one of the pioneering encoder-only Transformer models. Its uncased version, BERT-base uncased, signifies that the model does not differentiate between upper and lower-case letters, treating the text in a case-agnostic manner.
- RoBERTa (A Robustly Optimized BERT Pretraining Approach): As an enhancement to BERT, RoBERTabase [35] diverges from BERT in terms of training data and methodology, leading to more robust representations.
- 3. **ALBERT** (**A Lite BERT**): ALBERT-base v2 [36] is a more efficient variant of BERT, offering similar or even better performance using significantly fewer parameters. It achieves this by factorizing the large matrix into smaller matrices, thereby reducing redundancy.
- 4. **DeBERTa**: An advanced variant of BERT, DeBERTa improves upon BERT by utilizing a disentangled attention mechanism, which allows different parts of the model to focus on different types of information.
- 5. **PsychBERT**: Specifically tailored for psychological text analysis, PsychBERT embodies domain-specific knowledge making it especially suitable for tasks in the realm of psychology.

- ClinicalBERT: As the name suggests, ClinicalBERT
 is fine-tuned for clinical text, incorporating the nuances and lexicon commonly encountered in medical
 records and clinical narratives.
- MentalBERT: With a focus on mental health texts, MentalBERT captures patterns and contexts pertinent to mental well-being, disorders, treatments, and other related facets.

4.1.4. Decoder-only Transformer

On the other hand, the **decoder-only Transformer** primarily employs the decoder component of the Transformer architecture. Its strength lies in generating sequences, making it more apt for tasks like text generation, completion, and translation. Proposed by OpenAI in [37], GPT stands as one of the flagship decoder-only Transformers. It's pre-trained on vast text corpora to generate coherent and contextually relevant sequences of text.

It should be noted that while encoder-only models like BERT are designed for tasks that require understanding of text context, decoder-only models like GPT are optimized for generating coherent sequences. However, the boundaries between these applications have been blurred in recent advancements, with models being adapted for a wider array of tasks than their initial design intention.

Parameter Optimization: For consistency, we used the same experimental settings for all models and used 10 fold cross-validation. All results are reported as the average across all folds. We used the grid search optimization technique to optimize the parameters. To tune the number of layers (n), we empirically experimented with the values: learning rate $lr \in \{0.001, 0.005, 0.0001, 0.0005, 0.00001\}$ and optimization $O \in \{Adam, Adamax, AdamW\}$ with a batch-size of {8, 16, 32}. We used the base version pretrained language models (LMs) via HuggingFace⁶. We use optimized parameters for each baseline, and evaluate precision, recall, F1-score, and Accuracy. Varying lengths of posts are padded. We trained for 150 epochs with early stopping with a patience of 10 epochs. Thus, we set hyperparameter for our experiments with Transformer-based models as H = 200, O = Adam, learning rate = 1×10^{-5} , batch size 16, and 10 epochs. We further regularize LSTM and GRU with kernel regularization and bias regularization of 1×10^{-4} learning rate.

4.2. Quantitative Analysis

Table 5 displays the performance of different models on four wellness dimensions, namely Physical Aspect (PA), Social Aspect (SA), Emotional Aspect (EA), and Spiritual Aspect (SEA). From the table, it is evident that the GPT model and MentalBERT outperforms all the existing methods in terms of quantitative evaluation, indicating comparable results. While other models might have their respective strengths in various nuances, when it came to a holistic quantitative evaluation, these two models exhibited dominance. The GPT model and MentalBERT model leads the

⁶https://huggingface.co/models

pack, eclipsing the performance metrics of all other existing methodologies. A plausible explanation lies in the lexical characteristics associated with each dimension. The Social Aspect often involves direct terms like family, breakup, friends, etc., which might be easier for models to recognize. On the other hand, the Spiritual Aspect's lexical universe, containing words like motivation, happy, sad, etc., is broader and possibly harder to pinpoint to a single aspect.

The GPT model, built upon the Transformer architecture, is adept at capturing long-range contextual information in text. Its multiple attention mechanisms ensure that the model considers not just immediate neighboring words but also distant words to infer meaning. This becomes critical when determining context in intricate textual scenarios. MentalBERT, on the other hand, benefits from a nuanced fine-tuning process. While it borrows from the BERT architecture, its specific training on mental and wellnessoriented datasets primes it for excellence in tasks involving psychological and wellness dimensions. Its specialization makes it particularly effective in this domain. An interesting facet of the evaluation was the comparability of GPT and MentalBERT's results. While both were at the top of the leaderboard, their performance metrics were notably close. This suggests that while GPT's generalized training makes it a powerhouse in various NLP tasks, MentalBERT's specialized training ensures it's not far behind, at least in this domain. The neck-to-neck performance of these models presents an intriguing proposition for researchers and practitioners. It raises questions about the merits of broad generalized training versus specialized domain-specific training in mental healthcare domain. To this end, we further examine the attention over the text-spans that is used for decisionmaking by these models, respectively.

The Social Aspect (SA) pertains to interactions and relationships one has with their peers, family, and the broader community. Our analysis discerns that SA exhibits a unique linguistic signature. This is manifested in the form of specific words and phrases that are intricately woven into the social tapestry of human interactions. For instance: Family: this word encapsulates the primary social unit, indicating close bonds, familial ties, and inherent responsibilities. Its presence can immediately point towards topics related to kinship or family dynamics. Contrastingly, the Social-Emotional **Aspect (SEA)** encompasses a broader spectrum of emotions and sentiments that bridge both social interactions and individual emotional states. This dimension's vocabulary, while relevant, tends to be more generic and widespread across various contexts. For example, Motivation: While crucial to understanding one's drive and determination, motivation is a multifaceted term that can span numerous topics, not strictly limited to social-emotional contexts. Thus, we observe the highest and lowest scores for SA and SEA, respectively. This could be attributed to the fact that SA has a specific set of words associated with it, such as family, breakup, friends, ditched, partying, whereas SEA contains more general words, such as motivation, happy, sad, hate.

 Table 5

 Comparison of state-of-the-art methods. F-score, Precision, and Recall scores are averaged over 10 folds.

Method		PA			IVA			SA			SEA		Accuracy
	Р	R	F	Р	R	F	Р	R	F	Р	R	F	
LR	0.68	0.51	0.58	0.78	0.38	0.51	0.53	0.96	0.69	0.70	0.32	0.44	0.6009
RF	0.45	0.47	0.46	0.44	0.36	0.40	0.49	0.63	0.55	0.42	0.26	0.32	0.4604
LSTM	0.57	0.60	0.59	0.71	0.57	0.63	0.72	0.82	0.77	0.57	0.50	0.54	0.6591
GRU	0.43	0.73	0.54	0.80	0.45	0.57	0.82	0.77	0.79	0.52	0.34	0.41	0.6316
BERT	0.71	0.79	0.75	0.75	0.75	0.75	0.87	0.78	0.82	0.58	0.62	0.60	0.7464
RoBERTa	0.75	0.75	0.75	0.65	0.92	0.76	0.89	0.78	0.83	0.60	0.57	0.58	0.7480
ALBERT	0.75	0.73	0.74	0.70	0.77	0.73	0.82	0.83	0.82	0.61	0.57	0.59	0.7406
DeBERTa	0.80	0.76	0.78	0.82	0.76	0.79	0.80	0.88	0.84	0.66	0.62	0.64	0.7779
PsychBERT	0.76	0.73	0.74	0.75	0.80	0.78	0.82	0.86	0.84	0.65	0.58	0.61	0.7617
MentalBERT	0.77	0.75	0.76	0.79	0.81	0.80	0.83	0.88	0.85	0.68	0.61	0.64	0.7812
ClinicalBERT	0.74	0.77	0.75	0.73	0.75	0.74	0.83	0.86	0.84	0.68	0.57	0.62	0.7617
GPT-Ada	0.77	0.84	0.80	0.76	0.77	0.77	0.84	0.83	0.84	0.67	0.61	0.64	0.7779
GPT-Babbage	0.80	0.75	0.77	0.82	0.83	0.83	0.79	0.85	0.82	0.69	0.64	0.66	0.7795
GPT-Curie	0.78	0.78	0.78	0.77	0.81	0.79	0.81	0.84	0.82	0.67	0.59	0.63	0.7698
GPT-3 Davinci	0.79	0.78	0.79	0.82	0.84	0.83	0.80	0.85	0.83	0.68	0.59	0.63	0.7812

It was interesting to note that the identification of wellness dimensions is a challenging task, and the success of GPT can be attributed to its ability to interpret the contextual information present in social media text. However, there are some challenges that can make it difficult for pre-trained Transformers to accurately identify wellness dimensions in social media data. For example, social media data can be very informal and contain non-standard language, such as slang or abbreviations, which can be difficult for models to understand. Additionally, wellness dimensions can be highly subjective and context-dependent, which can make it challenging for models to accurately identify them. To ameliorate the issues posed by the idiosyncratic nature of social media text and the subjectivity in wellness dimensions, we advocate for a two-pronged approach: (i) Fine-tuning pretrained transformer models on specific datasets, like ours, to make them more attuned to the linguistic nuances and context, and (ii) Investing in research to design models with a robust capability to process informal language and discern context, enabling them to perform even better in such tasks.

4.2.1. Error Analysis

Given that identifying wellness dimensions in social media involves analyzing text data in the context of psychological concepts, the task is highly complex and specific to the field of social computing. Consequently, the challenges associated with this task are further compounded by the need to incorporate domain-specific psychology into the analysis. With this in mind, we have identified the following NLP-centered challenges that must be addressed to ensure accurate and effective analysis of social media data for wellness dimensions:

1. **Semantic Word Ambiguity**: Developing AI models for this task may result in semantic ambiguity during decision making. Consider a posts *P*₂ and *P*₃:

P₂: ...bad luck due to demons in my head...

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

Post P_2 affects SEA in its discussion about omen, while P_3 pertains to PA through physical injury or incapability causing mental disturbance. Although, the word *head* is used in both P_2 and P_3 , the semantic interpretations are different.

2. **Metaphors**: The cultural aspect of using metaphors is common practice in social media engagement [38, 39]. 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...

Although both P_1 and P_2 contain the content word drink, the expression drink myself to death is a metaphor suggesting IVA due to the financial risk of homelessness. On the contrary, P_2 points towards physical unfitness, i.e., PA.

3. Attention and Ambiguity: A surge in discourse and pragmatics suggests natural language understanding and low-level analysis to reduce ambiguity by identifying words more important than others, even when the same words are less important in other posts. For instance, consider the following post section:

 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...

P₄: I have 0 friends who would talk to me outside of work...

According to post P_0 , the author's true loss is the loss of an infant, affecting their SA. However, it also contains words such as *feeling shitty* and *work*, reflecting SEA and IVA, respectively. While words such as *work* are present but should not be emphasized in P_1 and P_4 , avoiding selection of *work* as potential text span depicting IVA. Instead, more attention is required on words such as *mom* and *friends* in P_1 and P_4 , respectively, thereby assigning it the SA category. Similarly, a given word *soul* must be emphasized as part of an adjective phrase in P_2 but as a noun in P_3 , thereby identifying it as IVA and SEA, respectively.

4.2.2. Experimental Inferences.

In our recent study, we observed that the baseline models, even those regarded as advanced, exhibited significant limitations, especially when considered for critical applications like mental health assessment. Our rigorous evaluations indicated that, contrary to popular belief, even the state-of-the-art Transformer models are far from perfect, often making errors that could be deemed as careless in high-stakes environments. A pertinent concern that arose was the issue of overfitting. When these models are trained and subsequently evaluated on in-domain data drawn from a singular distribution, they tend to become exceedingly attuned to the peculiarities of that data. As a result, their performance may degrade considerably when introduced to out-of-domain data or data that showcases different attributes. A prime example is how a model trained on data from one social media platform might flounder when faced with data from another platform. This could be attributed to a myriad of factors, ranging from differences in user health behaviors, variances in socio-demographic indicators, or even the unique linguistic idiosyncrasies that users on different platforms exhibit. These intricacies necessitate a deep domain knowledge, which our WELLXPLAIN dataset strives to encapsulate. Thus, while the promise of AI in mental health is undeniable, our findings reiterate the importance of exercising caution, especially when the stakes are high.

4.2.3. Discussion.

Wellness dimensions, by their very nature, are intricately interwoven. They often don't manifest as distinct, separate entities, but rather as complex intersections of various facets of an individual's well-being. A significant chunk of the discrepancies is traceable to the overlap between the Physical Aspect (PA) and the combined Spiritual and Emotional Aspect (SEA). The reason for this is straightforward: many activities or events impinge on multiple wellness dimensions simultaneously. Taking the example of crying, it's an action that exacts a physical toll (physical fatigue, dehydration, etc.) even as it signals a deep emotional upheaval. Thus,

determining whether to categorize it under PA or SEA is intrinsically challenging. Social media posts, often reflective of real-life experiences, are rarely one-dimensional. A single post might touch upon various wellness dimensions, further complicating the annotation process. For instance, a post discussing a rigorous yoga session could touch upon the physical exertion involved (PA) and the ensuing mental tranquility (SEA). The inherent subjectivity and variability of human emotions mean that many posts don't fit neatly into predefined categories. Ambiguity in textual content makes it hard to pin down the exact wellness dimension it pertains to. Reading between the lines, understanding subtext, or gauging the unsaid is a skill that varies among annotators. This variability is a fertile ground for manual disagreements.

4.3. Qualitative Analysis

To begin our investigation into monitoring the progression of mental disturbance, we concentrate on the wellness concept extraction through relevant text-spans. This preliminary study is a crucial step towards understanding the underlying mechanisms that lead to severe mental illnesses such as depression and self-harm by identifying wellness concept in a given text. Future work will aim to incorporate additional factors to gain a more comprehensive understanding positivity and negativity of wellness concepts in Reddit posts. In this section, we examine the reliability and explainability of the traditional multi-class classifiers.

Reliability/ Explainability Analysis. In our endeavor to assess the clarity and understandability of model predictions, we juxtaposed the explanations derived from ground truth with those garnered through the LIME method, as delineated in [40]. Our experimental setup involved extracting pivotal terms from a sample of 100 randomly selected data points, employing both recurrent models and Transformer architectures. LIME's strength lies in its ability to identify and highlight those specific words within the data that most potently sway the classifier's verdict.

In our research, the analytical results, captured comprehensively in Table 6, incorporated two renowned metrics: ROUGE and BLEU. These metrics are universally recognized benchmarks in the domain of natural language processing, particularly when one seeks to evaluate the quality and relevance of generated text against a predetermined reference or gold standard. ROUGE, which stands for Recall-Oriented Understudy for Gisting Evaluation, primarily assesses the recall rate of generated content, spotlighting how many of the reference's components (words, phrases, etc.) were accurately captured in the produced text. On the other hand, BLEU (Bilingual Evaluation Understudy) emphasizes precision, ensuring that the generated content's components are indeed present in the reference text. Together, these metrics offer a holistic evaluation, gauging both the comprehensiveness and accuracy of generated explanations relative to a benchmark. Through the simultaneous application of ROUGE and BLEU, we ensured a thorough and robust assessment of the explanations' fidelity to the reference standard.

Table 6
LIME's explanation scores using ROUGE-1.

Method	F-score	Precision	Recall	Bleu-1	Bleu-2
BERT	0.4668	0.3970	0.7076	0.3687	0.1311
RoBERTa	0.4170	0.3197	0.8525	0.3651	0.1302
ALBERT	0.4660	0.3954	0.6999	0.3682	0.1342
DeBERTa	0.4564	0.3886	0.6852	0.3614	0.1342
PsychBERT	0.4581	0.3890	0.6879	0.3647	0.1383
MentalBERT	0.4866	0.4095	0.7463	0.3827	0.1456
ClinicalBERT	0.4691	0.3985	0.7105	0.3720	0.1401
GPT-Ada	0.5002	0.5548	0.5543	0.4078	0.3485
GPT-Babbage	0.5267	0.5944	0.5761	0.4226	0.3591
GPT-Curie	0.5213	0.5874	0.5722	0.4240	0.3582
GPT-Davinci	0.5335	0.5930	0.5916	0.4361	0.3758

Our experimental outcomes presented insightful revelations about the performance of various language models. Notably, GPT-3 ascended as the frontrunner, demonstrating unparalleled precision, recall, and F-score metrics, illustrating its prowess in accurately identifying and spotlighting critical words within generated sequences (See table 6). Following closely was the performance exhibited by the MentalBERT model. Unlike GPT-3, which gauges impactful words through its generated sequences, MentalBERT employs the LIME methodology, with a specific emphasis on the attention mechanism to earmark salient words.

It's pertinent to mention that while GPT-3 and Mental-BERT both achieved commendable results, drawing a direct comparison between the two may not be straightforward, given their differing methodologies and underlying architectures. Notwithstanding, the distinctions in their results were conspicuous, thereby underscoring the respective strengths and nuances of each approach.

A key observation that further emerged from our study was the superior reliability offered by expansive generative language models like GPT-3. Despite the tailor-made finetuning that models like MentalBERT undergo for domain-specific tasks, the expansive training and inherent capabilities of models like GPT-3 enable them to outshine in various tasks. This observation potentially accentuates the evolving dynamics of AI models, wherein broad-spectrum models, thanks to their extensive training and diversified datasets, might exhibit more robust and reliable performances, even when juxtaposed against domain-specialized transformers.

Practical Implications. We suggest longitudinal studies for fair and accountable practices of analyzing the emotional spectrum of users' historical social media posts [41]. The ongoing research aimed at identifying changes in moments from users' longitudinal social media posts [42] should benefit from our WELLNESS DIMENSIONS dataset. 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.

Post T1 illustrates the confused state of the social media user. Movement of thoughts towards politics $(T2 \rightarrow T3)$, relationships (T4) and finally towards suicidal intention (T5) illustrates the tremendous impact of users' experiences on different wellness dimensions over time. Henceforth, our dataset shall support and compliment other studies [43, 44] as an intrinsic classification task.

Limitations and Ethical Consideration We acknowledge that our work is subjective in nature and thus, interpretation about wellness dimensions in a given post may vary from person to person. Clearly, machine learning predictions are unable to replace professional mental health diagnostic let alone counseling and therapy. As shown in our evaluation, their accuracy and trustworthiness remain insufficient for such purposes [45]. Rather, we are hoping to engender further research on how to recognize early warning signs that may otherwise go unnoticed and neglected. For this, it is important to obtain consent for all relevant use cases. It is also important for human professionals to carefully assess model predictions before any pertinent action is taken. Our work promotes explainability to facilitate human validation [46]. However, even here it must be noted that explanations can be misleading and that human validators need to very carefully review the entire post context.

We emphasize the importance of preserving privacy due to the sensitive nature of social media data [10, 46]. To ensure our accountability, we will provide appropriate protections for sensitive data, and there will be no linkage of the dataset to other sites that could jeopardize user anonymity [47]. We further acknowledge that suggesting professional help based on a given social media post is a starting point for this study. In our work, *explainability*

refers to the text spans that appears as indicators of wellness dimensions in a given text.

5. Conclusion

In this work, we present WELLXPLAIN, a newly constructed dataset for wellness concept extraction and classification to facilitate the future research in this direction. This work is based on a new task definition of clinical concept extraction and classification with a carefully designed annotation scheme, including perplexity guidelines. Furthermore, we solicit class annotations and text spans for explainability purposes. In addition, we present baseline experiments conducted on a diverse range of methods. The contribution of this work derives from the potential for tackling different use cases of wellness dimensions at a deeper, interpretable level. We endeavor to disseminate this position widely in the research community and urge researchers to develop richer, explainable models for inferring mental illness on social media. We keep data augmentation and few-shot learning approaches as open research directions to develop efficient AI models. With this work, we seek to promote investigating mental health diagnostics that operate at a deeper, interpretable level and hope that future work benefits from our data. Furthermore, by considering the different dimensions of wellness, Social Determinants of Health 2030 can create policies that address the various aspects of wellbeing, leading to a more holistic evaluation [48].

Acknowledgements

We would like to convey our heartfelt thanks to our postgraduate student annotators, Ritika Bhardwaj, Astha Jain, and Amrit Chadha, for their meticulous contributions to the annotation process. Our deep gratitude goes to Veena Krishnan, an esteemed clinical psychologist, and Ruchi Joshi, a dedicated rehabilitation counselor, for their steadfast support and invaluable insights throughout the duration of this project. Furthermore, our warm appreciation is directed towards Prof. Sunghwan Sohn for his unwavering mentorship and guidance throughout our journey.

References

- UN, Transforming our world: The agenda 2030 for sustainable development (2015).
- [2] M. F. Garnett, S. C. Curtin, D. M. Stone, Suicide mortality in the united states, 2000–2020, NCHS data brief (433) (2022) 1–8.
- [3] E. L Belfort, E. Mezzacappa, K. Ginnis, Similarities and differences among adolescents who communicate suicidality to others via electronic versus other means: a pilot study, Adolescent Psychiatry 2 (3) (2012) 258–262.
- [4] P. Resnik, A. Foreman, M. Kuchuk, K. Musacchio Schafer, B. Pinkham, Naturally occurring language as a source of evidence in suicide prevention, Suicide and Life-Threatening Behavior 51 (1) (2021) 88–96.
- [5] C. Berryman, C. J. Ferguson, C. Negy, Social media use and mental health among young adults, Psychiatric quarterly 89 (2) (2018) 307– 314.

- [6] A. Pourmand, J. Roberson, A. Caggiula, N. Monsalve, M. Rahimi, V. Torres-Llenza, Social media and suicide: a review of technologybased epidemiology and risk assessment, Telemedicine and e-Health 25 (10) (2019) 880–888.
- [7] B. O'dea, M. E. Larsen, P. J. Batterham, A. L. Calear, H. Christensen, A linguistic analysis of suicide-related twitter posts., Crisis: The Journal of Crisis Intervention and Suicide Prevention 38 (5) (2017) 319.
- [8] M. A. Franco-Martín, J. L. Muñoz-Sánchez, B. Sainz-de Abajo, G. Castillo-Sánchez, S. Hamrioui, I. de La Torre-Díez, A systematic literature review of technologies for suicidal behavior prevention, Journal of medical systems 42 (2018) 1–7.
- [9] J. Du, Y. Zhang, J. Luo, Y. Jia, Q. Wei, C. Tao, H. Xu, Extracting psychiatric stressors for suicide from social media using deep learning, BMC medical informatics and decision making 18 (2018) 77–87.
- [10] K. Harrigian, C. Aguirre, M. Dredze, On the state of social media data for mental health research, NAACL HLT 2021 (2021) 15.
- [11] S. Roy, S. Pfohl, G. A. Tadesse, L. Oala, F. Falck, Y. Zhou, L. Shen, G. Zamzmi, P. Mugambi, A. Zirikly, et al., Machine learning for health (ml4h) 2021, in: Machine Learning for Health, PMLR, 2021, pp. 1– 12.
- [12] M. Choudhury, D. Chatterjee, A. Mukherjee, Global topology of word co-occurrence networks: Beyond the two-regime power-law, in: Coling 2010: Posters, 2010, pp. 162–170.
- [13] M. Gaur, A. Alambo, J. P. Sain, U. Kursuncu, K. Thirunarayan, R. Kavuluru, A. Sheth, R. Welton, J. Pathak, Knowledge-aware assessment of severity of suicide risk for early intervention, in: The World Wide Web Conference, 2019, pp. 514–525.
- [14] T. E. Joiner Jr, K. A. Van Orden, T. K. Witte, M. D. Rudd, The interpersonal theory of suicide: Guidance for working with suicidal clients., American Psychological Association, 2009.
- [15] H. L. Dunn, High-level wellness for man and society, American journal of public health and the nations health 49 (6) (1959) 786–792.
- [16] G. Coppersmith, M. Dredze, C. Harman, K. Hollingshead, M. 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, 2015, pp. 31–39.
- [17] G. Shen, J. Jia, L. Nie, F. Feng, C. Zhang, T. Hu, T.-S. Chua, W. Zhu, Depression detection via harvesting social media: A multimodal dictionary learning solution., in: IJCAI, 2017, pp. 3838–3844.
- [18] A. Yates, A. Cohan, N. Goharian, Depression and self-harm risk assessment in online forums, in: EMNLP, 2017.
- [19] A. Cohan, B. Desmet, A. Yates, L. Soldaini, S. MacAvaney, N. Goharian, Smhd: a large-scale resource for exploring online language usage for multiple mental health conditions, in: 27th International Conference on Computational Linguistics, ACL, 2018, pp. 1485–1497.
- [20] D. E. Losada, F. Crestani, J. Parapar, Overview of erisk: early risk prediction on the internet, in: International conference of the crosslanguage evaluation forum for european languages, Springer, 2018, pp. 343–361.
- [21] L. Cao, H. Zhang, L. Feng, Z. Wei, X. Wang, N. Li, X. 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), 2019, pp. 1718–1728.
- [22] E. Turcan, K. McKeown, Dreaddit: A reddit dataset for stress analysis in social media, arXiv preprint arXiv:1911.00133 (2019).
- [23] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, S. Ravi, Goemotions: A dataset of fine-grained emotions, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 4040–4054.
- [24] H.-C. Shing, P. Resnik, D. W. Oard, A prioritization model for suicidality risk assessment, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8124–8137.

- [25] A. Haque, V. Reddi, T. Giallanza, Deep learning for suicide and depression identification with unsupervised label correction, in: International Conference on Artificial Neural Networks, Springer, 2021, pp. 436–447.
- [26] 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.
- [27] M. Garg, Mental health analysis in social media posts: A survey, Archives of Computational Methods in Engineering (2023) 1–24.
- [28] A. Tsakalidis, F. Nanni, A. Hills, J. Chim, J. Song, M. Liakata, Identifying moments of change from longitudinal user text, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 4647–4660.
- [29] C. T. Kitko, Dimensions of wellness and the health matters program at penn state, Home Health Care Management & Practice 13 (4) (2001) 308–311.
- [30] K. A. Strout, E. P. Howard, The six dimensions of wellness and cognition in aging adults, Journal of Holistic Nursing 30 (3) (2012) 195–204.
- [31] T. Mitra, E. Gilbert, Credbank: A large-scale social media corpus with associated credibility annotations, in: Proceedings of the international AAAI conference on web and social media, Vol. 9, 2015, pp. 258–267.
- [32] J. L. Fleiss, Measuring nominal scale agreement among many raters., Psychological bulletin 76 (5) (1971) 378.
- [33] M. Zulqarnain, R. Ghazali, Y. M. M. Hassim, M. Rehan, Text classification based on gated recurrent unit combines with support vector machine, International Journal of Electrical and Computer Engineering 10 (4) (2020) 3734.
- [34] J. D. M.-W. C. Kenton, L. K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL-HLT, 2019, pp. 4171–4186.
- [35] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).
- [36] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, Albert: A lite bert for self-supervised learning of language representations, in: International Conference on Learning Representations, 2019.
- [37] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al., Improving language understanding by generative pre-training (2018).
- [38] C. A. Bail, Cultural carrying capacity: Organ donation advocacy, discursive framing, and social media engagement, Social Science & Medicine 165 (2016) 280–288.
- [39] M. Chmielecki, et al., Conceptual negotiation metaphors across cultures-research findings from poland, china, the united states and great britain, Journal of Intercultural Management 5 (3) (2013) 103– 118.
- [40] 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.
- [41] R. Sawhney, H. Joshi, S. Gandhi, R. R. Shah, Towards ordinal suicide ideation detection on social media, in: Proceedings of the 14th ACM International Conference on Web Search and Data Mining, 2021, pp. 22–30.
- [42] A. Tsakalidis, J. Chim, I. M. Bilal, A. Zirikly, D. Atzil-Slonim, F. Nanni, P. Resnik, M. Gaur, K. Roy, B. 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, 2022, pp. 184– 198.
- [43] T. Saha, V. Gakhreja, A. S. Das, S. Chakraborty, S. 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, 2022, pp. 2650–2656.

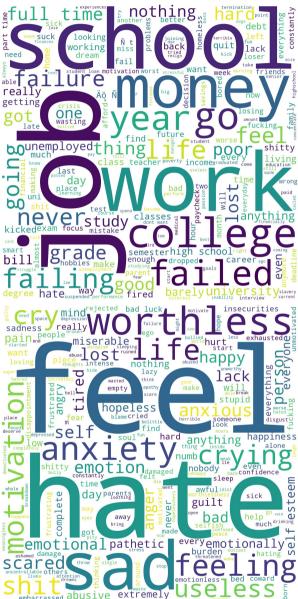
- [44] M. Gaur, V. Aribandi, A. Alambo, U. Kursuncu, K. Thirunarayan, J. Beich, J. Pathak, A. Sheth, Characterization of time-variant and time-invariant assessment of suicidality on reddit using c-ssrs, PloS one 16 (5) (2021) e0250448.
- [45] J. Nicholas, S. Onie, M. E. Larsen, Ethics and privacy in social media research for mental health, Current Psychiatry Reports 22 (12) (2020) 1–7
- [46] S. Chancellor, M. L. Birnbaum, E. D. Caine, V. M. Silenzio, M. De Choudhury, A taxonomy of ethical tensions in inferring mental health states from social media, in: Proceedings of the conference on fairness, accountability, and transparency, 2019, pp. 79–88.
- [47] H.-C. Shing, S. Nair, A. Zirikly, M. Friedenberg, H. Daumé III, P. Resnik, Expert, crowdsourced, and machine assessment of suicide risk via online postings, in: Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic, 2018, pp. 25–36.
- [48] C. A. Gómez, D. V. Kleinman, N. Pronk, G. L. W. Gordon, E. Ochiai, C. Blakey, A. Johnson, K. H. Brewer, Practice full report: Addressing health equity and social determinants of health through healthy people 2030, Journal of Public Health Management and Practice 27 (6) (2021) S249.

Appendix

A. Word Frequency in Explanations

Word clouds for explanations of different wellness dimensions are shown in Figure 2.





abusive

Figure 2: Top: Word cloud for Physical Aspect (Left) and Intellectual and Vocational Aspect (Right);// Bottom: Word cloud for Social Aspect (Left) and Spiritual and Emotional Aspect (Right)