Vanishing Boundaries: A Unifying Account of Multidimensional Emotion Dynamics and Alterations in Depression

Ana-Maria Bucur[1,2], Tahmineh A. Koosha[3], Adrian Cosma[4], Lucie Flek[5, 6*], Sharmili Edwin Thanarajah[7, 8], Felix Bernhard[3], Paolo Rosso[2,9*], Hamidreza Jamalabadi[3*]

[1]Interdisciplinary School of Doctoral Studies, University of Bucharest, Romania.
[2]PRHLT Research Center, Universitat Politècnica de València, Spain.
[3]Department of Psychiatry and Psychotherapy, Philipps University of Marburg, Germany.
[4]Computer Science Department, National University of Science and Technology Politehnica Bucharest, Romania.
[5]Conversational AI and Social Analytics (CAISA) Lab, Bonn-Aachen International Center for Information Technology (b-it), University of Bonn, Germany.
[6]The Lamarr Institute for Machine Learning and Artificial Intelligence, Germany.
[7]Department of Psychiatry, Psychosomatic Medicine and Psychotherapy, University Hospital, , Goethe University, Frankfurt, Germany.
[8]Max Planck Institute for Metabolism Research, Cologne, Spain.
[9]ValgrAI Valencian Graduate School and Research Network of Artificial Intelligence, Spain.

Contributing authors: ana-maria.bucur@drd.unibuc.ro; tahmineh.koosha@uni-marburg.de; ioan_adrian.cosma@upb.ro; flek@bit.uni-bonn.de; edwinthanarajah@med.uni-frankfurt.de; felix.bernhard@staff.uni-marburg.de; prosso@dsic.upv.es; hamidreza.jamalabadi@uni-marburg.de;

Abstract

Emotions are fundamentally integral to shaping the order and disorders in human lives. Yet, a principled, quantitative framework explaining emotional dynamics and their alteration in mental disorders has been elusive. This challenge arises from the complex and multidimensional nature of emotions but also, at least partially, due to a shortage of large longitudinal measurements and the use of generative mathematical models, leading to a spectrum of partially contrasting theories. Our study seeks to overcome these challenges by employing dynamic systems theory, the mathematical study of complex systems. We apply this approach to a dataset containing over 400,000 texts from more than 1,600 individuals, with half reporting a diagnosis of Major Depressive Disorder (MDD), collected from X (previously Twitter). We examined the emotional experiences and significant life events described in these texts across one month, which we extracted using state-of-the-art natural language processing. The key result of our research is the discovery of emotion-specific dynamics – the unique ways in which different emotions maintain their influence over time. In individuals diagnosed with MDD, we observed a 'blunting' of the inter-emotional dynamics; this is not merely a dulling of emotions, but rather a vanishing of the boundaries between them. Our findings thus challenge and unify traditional views in this area: if viewed in isolation, this blending of emotions could be misinterpreted as augmented negative, blunted positive emotions, and as Emotional Context Insensitivity (ECI). Our study therefore offers a principled mathematical understanding of emotion dynamics and their alterations within mental disorders, potentially leading to more effective therapeutic interventions.

Keywords: Emotion dynamics, Depression, Natural language processing, Dynamic systems theory

1 Introduction

Emotions are central to our human experience and play a crucial role in our mental well-being. Growing evidence shows that emotions are dynamic, constantly evolving in response to our interactions with the environment [1]. Traditionally though, emotions have been studied as either stable traits or as transient states that quite statically activate or deactivate in response to particular events. Yet, recent research is shifting towards a more dynamic understanding of emotions [2]. This new focus extends beyond the simplistic categorization of emotions as positive or negative. It adopts a dimensional approach that aligns more closely with the neural foundations of emotional processes [3], highlighting the significance of comprehending the development, evolution, and interaction of emotions.

Key concepts in the study of emotion as dynamic constructs include principles of contingency, inertia, and regulation [4, 5]. These principles highlight the multidimensional nature of emotions, their interactions over time (contingency), their tendency to persist despite changing circumstances (inertia), and their susceptibility to internal and external influences (regulation). From a computational perspective, the

literature introduces multiple measures to estimate the dynamic properties of emotions, yet these measures often rely on simple statistical heuristics, primarily based on mean, variance, and correlations (see [2, 6] for a list and mathematical definitions). While these metrics provide valuable insights into the probabilistic characteristics of emotions, they fall short of capturing their temporal dynamics e.g., in the context of pathological conditions such as Major Depressive Disorder (MDD) [6]. For instance, there is an active ongoing debate on how shifts in emotional dynamics manifest in mental disorders, with three partially contrasting theories being proposed: positive attenuation (a weakened response to positive stimuli), negative potentiation (enhanced response to negative stimuli), and Emotional Context Insensitivity (ECI; a diminished response to both positive and negative emotions) [7–9].

In this paper, we aim to tackle these complexities by building a mathematical model of how emotions evolve, interact, and change in response to life events [10]. At the core of our approach is the Dynamic Systems Theory (DST), an analytical framework for examining complex systems over time [11] which has recently proven invaluable to understanding human behavior, offering insights into phenomena such as the progression of depressive symptoms and strongly accurate prediction of the response to psychological interventions[12–14]. In DST, the evolution of emotions can be mathematically quantified through differential or time-recursive equations. Typically, these equations are linear, offering the advantage of being both highly understandable and amenable to efficient, data-driven estimation techniques, facilitating a direct linkage to the underlying mechanisms of emotion dynamics [15]. Concrete, we utilize three key parameters derived from DST (detailed in the Methods section). The first parameter, *emotional reactivity*, measures the expected variation in the emotional landscape triggered by a change in a specific emotion. This parameter reflects the interconnectedness of emotional experiences. It extends traditional concepts of emotional reactivity into a more nuanced, mathematically grounded framework. Second, we calculate the *time constant* for each emotional variable, which mathematically denotes the rate at which an emotion evolves over time. Our method independently assesses the rate of change for each emotion, offering a more precise understanding of emotional evolution free from the biases inherent in other approaches [16–18]. Third, we introduce *emotion controllability*, a novel metric to evaluate the influence of external life events on emotional dynamics. Unlike existing literature, which often relies on pre-post measures, our method quantitatively assesses how life events affect the trajectory of emotional changes (see Figure 1).

Towards this end, we analyze these three parameters to study an extensive dataset comprising over 400,000 texts from more than 1,600 individuals, half of them self-reporting MDD diagnosis, collected over an average duration of one month [19]. We utilize EmoRoBERTa [20], a state-of-the-art transformer-based model, to estimate the emotions expressed in these texts. This process estimates 28 distinct emotions, providing a comprehensive emotional profile for each individual across the one-month period, with updates roughly nine times daily (see Figure 1 and Methods for details). Further, we apply a separate machine learning model to identify and categorize life events (more specifically, "happy moments") mentioned in the texts [21, 22]. These events are classified as positive life events, such as expressions of joy, contentment,

or satisfaction found in natural language. Examples include statements like "I went for a run in the neighborhood. I enjoyed the perfect weather" and "We booked our beach vacation for May of this year." In light of existing research on the patterns of emotional and environmental disengagement associated with depression, our study followed three hypotheses: 1) We expect to find heightened reactivity to negative emotions, reflecting an intensified response to adverse emotional stimuli among individuals with depression. 2) We anticipate a noticeable reduction in the time constants of emotions for those diagnosed with depression. This suggests a slower pace of emotional change, particularly for negatively charged emotions, indicating a prolonged emotional response. 3) We hypothesize that positive life events will have a diminished effect on individuals with depression, suggesting a decreased ability to experience uplifts from happy occurrences.

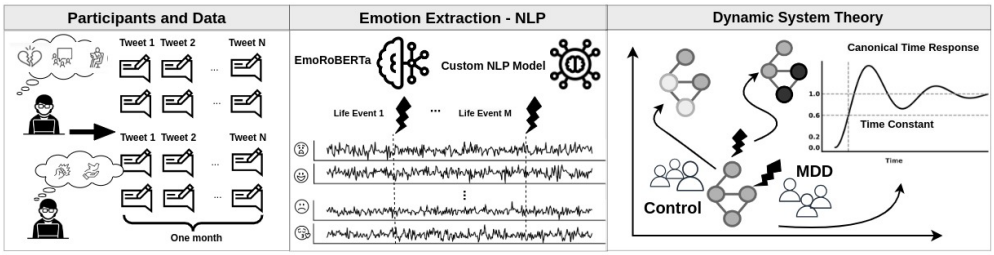

Fig. 1: **Overview of Research Methodology**. This study develops a mathematical model to analyze emotional dynamics, the interaction between different emotions, and how these are influenced by life events and mental health, specifically focusing on Major Depressive Disorder (MDD). Utilizing over 400,000 text entries from over 1,600 participants—half reporting a diagnosis of MDD — this work leverages EmoRoBERTa for emotion assessment and a customized machine learning model for categorizing significant life events. Employing Dynamic Systems Theory and, in particular, control theory, we build generative models of multidimensional emotion evolution. This approach enables us to accurately determine the canonical response of the emotion trajectories to both external factors, such as life events, and internal shifts among different emotions.

2 Results

To model the dynamics of emotions, we began by analyzing the emotional expressions contained within the texts of all 1,690 participants. This was accomplished using a transformer-based EmoRoBERTa model [20] (see Methods for details), which is capable of identifying a spectrum of 28 distinct emotions (see Figure 2). The model outputs an estimation of the probability for each identified emotion within the texts at various time points. We systematically extracted this data for each participant over a 30-day period, averaging nine emotional assessments per day. In the next step, we examined

the emotion dynamics in two measures: First, we determined the emotional time constant for each emotion, as illustrated in Figure 2. The time constant is a measure that captures the average standardized time it takes for an emotion to either intensify or diminish following a life event or its conclusion and is thus closely related to the concept of emotional inertia [17]. We found that "love" had the highest time constant, indicating its long-lasting impact over time. Conversely, "embarrassment" had the lowest time constant, suggesting it is a rapidly diminishing emotional response. A comparison with depressed subjects revealed a significant effect: The absolute value of the time constants for all but four emotions in the depressed group were lower (p-value < 0.001; binomial test). On average, we observed a 30% reduction in the size of time constants for emotions. This indicates that in individuals with depression, the duration for emotional responses to fluctuate—either to escalate or to recede—is significantly different from those without depression, suggesting a 'blunted' emotional landscape, where emotions can either fade away faster or linger beyond the levels expected in the healthy population, consistent with both accounts of emotional blunting and enhanced negative emotions. Specifically, it appears that various emotions in depressed individuals tend to converge towards a generic, unified duration, losing their distinct temporal patterns. Notably, this blunting effect is most evident in emotions like "nervousness" and "embarrassment," which are key factors considered in depression assessments such as the Beck Depression Inventory (BDI) [23]. Other emotions like "sadness," "remorse," "relief," "love," "fear," and "disgust," integral to depression diagnostic criteria, also showed significant blunting. These findings suggest a characteristic alteration in the temporal dynamics of emotions in depression, deviating from the typical emotional response patterns seen in healthy individuals.

In our second analysis, we examined the interconnectedness of emotional responses, exploring how a shift in one emotion impacts others using a measure we call emotional reactivity. We quantified emotional reactivity by calculating the average total effect that any given emotional change has on the spectrum of possible changes in all other emotions (using a metric called average controllability, see Methods for details). Emotions with higher reactivity are, therefore, more influential in determining the multidimensional temporal evolution of emotions [24]. Our findings, illustrated in Figure **??**, indicated that "grief" and "fear" possess the highest emotion reactivity, meaning they significantly affect other emotions. On the other hand, "neutral" emotions, as expected, exhibit the least emotional reactivity, followed closely by "approval". When comparing individuals with depression to healthy individuals, we observed a notable pattern: a general blunting of cross-emotional reactivity differences. Except for three specific emotions ("admiration," "amusement," and "curiosity"), the size of emotional reactivity was always higher in healthy individuals than in those with depression (p-value $<$ 1e-6, binomial test). On average, we observed a 43% reduction in the size of emotion reactivity for emotions. This pattern aligns with existing theories of emotion regulation in depression, particularly with the concept of context insensitivity. However, it also provides a clearer understanding of why observations vary across studies. Significantly, the most substantial blunting was observed for "neutral" emotions. This finding is consistent with previous research indicating biases in how individuals with

depression assess neutral emotions [25]. Blunting was also prominent in emotions like "relief" and "pride," which are integral to depression diagnostic criteria.

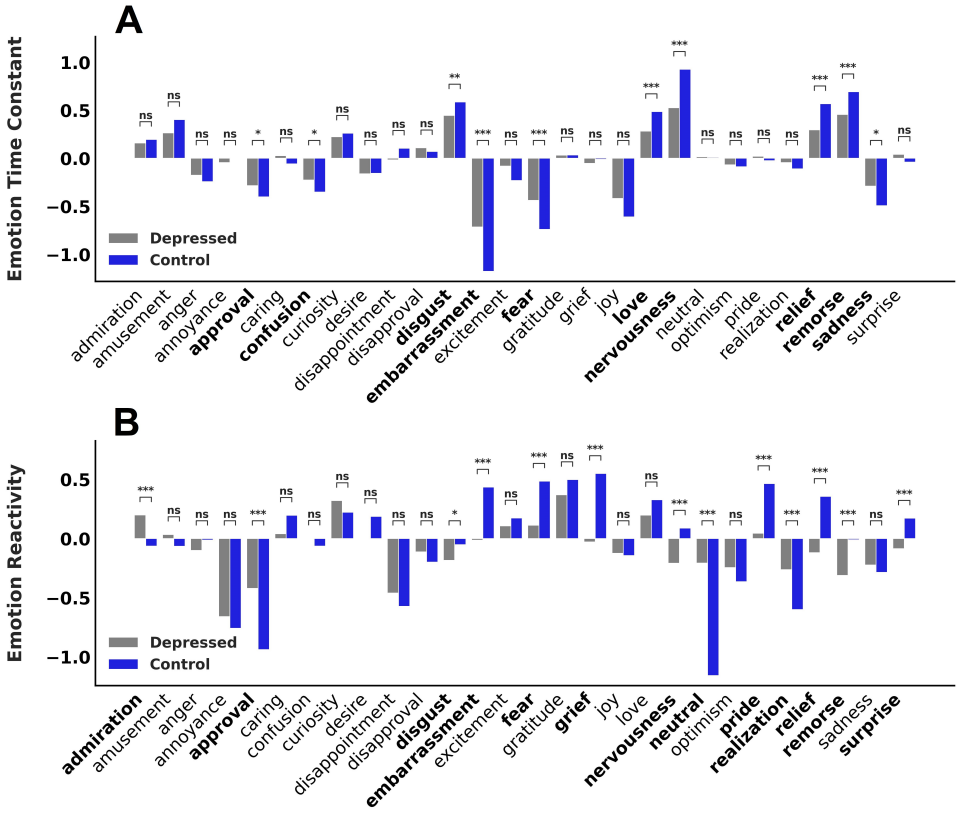

Fig. 2: **Dynamics of Emotions in Individuals With and Without Major Depressive Disorder (MDD).** (A) Emotional Time Constant: Illustrates the average time it takes for emotions to intensify or lessen post-event, serving as a proxy for emotional inertia and stability. (B) Emotional Reactivity: This quantifies the extent to which a single emotional shift can influence the overall emotional matrix, utilizing a metric called average controllability (see Methods for details). Higher reactivity denotes a significant impact on the unfolding of emotional dimensions. Significance levels are based on Bonferroni corrections and are indicated as follows: *ns* ($0.05 < p \leq 1$), $*$ ($0.01 < p \leq 0.05$), $**$ ($0.001 < p \leq 0.001$), $***$ ($1. \times 10^{-4} < p \leq 1 \times 10^{-3}$), and $****$ ($p \leq 1 \times 10^{-4}$).

To explore whether the emotional blunting in MDD extended to the reactivity to life events, we identified life events representing happy moments [26] in our dataset (detailed in the Methods section). Of note, prior research indicates that methods

focused solely on classifying sentiments or emotions are inadequate for identifying expressions of happiness [27], as they can be conveyed in a variety of linguistic tones, not just positive ones [28]. "Happy moments" are expressions in natural language reflecting a subjective experience of joy, contentment, or satisfaction. Our approach utilized a comprehensive dataset of 100,000 "happy moments" collected from over 10,000 contributors [26]. We then developed a specialized deep learning model (all codes publicly available) designed to detect and isolate instances of happy moments in text, enhancing the precision of our emotional analysis (for a detailed explanation, see the Methods). The extracted happy moments covered a wide range of topics, such as relationships (e.g., "girlfriend said she loves me."), family (e.g., "I'm celebrating my nephew's birthday!", "Lunch date with my sister"), life (e.g., "Finally getting my life back together, after the downhill I had.", "My birthday is on Monday!"), food (e.g., "I'm scooping ice cream at a diner near my town!"), home decoration (e.g., "We have ordered a new bookcase - excited to be able to keep more books!"). We found happy moments in approximately 3.7% of the texts from patients with MDD and 2.4% from healthy individuals.

Based on this data, we tested if the impact of life events on individuals with depression was less pronounced compared to their effect on healthy individuals. This inquiry aligns with our earlier observations of blurring emotional distinctions in depression, marked by shortened emotional response times (inertia) and decreased emotional reactivity. We hypothesized that, against this backdrop of diminished emotional differentiation, isolated happy events would exert a weaker influence on the emotional trajectories of individuals with depression than on those of healthy individuals. [29]. To investigate this hypothesis, we applied principles from Dynamic Systems Theory, specifically using the concept of control energy. Control energy measures the effort required to alter the collective state of a system. In the context of our study, this meant evaluating the extent to which various events could alter emotional trajectories. Our findings, depicted in Figure 3, show that for individuals with depression, happy moments have a reduced capacity to influence emotional states. Our findings rely on analyzing control energy in complex systems through two critical measures: the minimum and average eigenvalues of the Gramian matrix [30] (for more details, see the Methods section). The marked disparity we observed underscores the altered emotional responsiveness in people with depression, particularly highlighting the phenomenon of emotional blunting. This phenomenon is marked by diminished reactions to emotional triggers in depressed individuals.

To rigorously compare these metrics, we faced the challenge of non-normal data distribution. To address this, we utilized the non-parametric Mann-Whitney U test. This statistical approach revealed a significant difference (with a p-value less than 0.005) between the control group and the group with MDD.

3 Discussion

By conceptualizing human emotions as observable elements of a complex dynamic system, we can apply the foundational principles of dynamic systems theory to decipher the temporal evolution of emotions in individuals, both with and without depression,

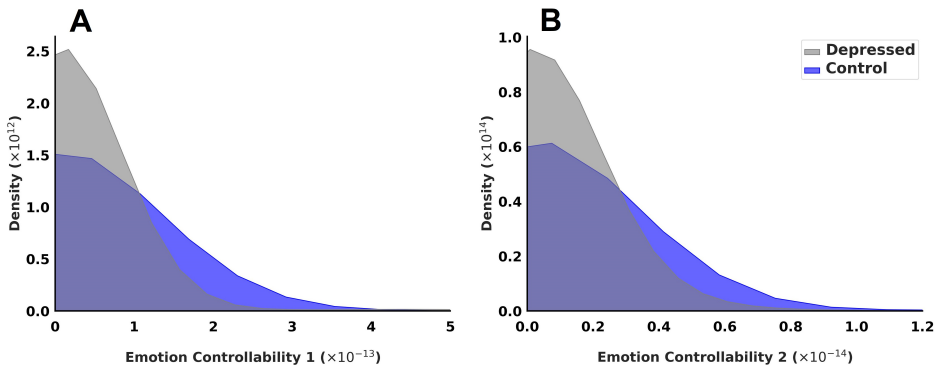

Fig. 3: **Effects of Life Events on Emotional Development: The Role of Depression.** This figure highlights how life events impact emotional evolution, especially how this impact is altered by depression. It demonstrates that individuals with depression experience a lower control energy from life events, requiring significantly stronger events to match the emotional effect observed in healthy counterparts. (A) Emotion Controllability 1: Estimated through the average eigenvalues of the Gramian matrices for individuals with and without depression. (B) Emotion Controllability 2: Estimated through the minimum eigenvalues of the Gramian matrices for individuals with and without depression. Given the non-normal distribution of the two metrics, we used the non-parametric Mann-Whitney U test.

and in response to life events. The core findings of our research highlight the temporal nuances of emotions, manifesting through their specific time constants and patterns of joint activation (i.e., emotional reactivity). These elements collectively delineate a rich, multidimensional emotional landscape. In the context of depression, we observe a marked shift in this landscape: it becomes more homogenized, with the distinctions between different emotions becoming increasingly indistinct. This finding offers deep insights that resonate with, and yet significantly broaden, the traditional understanding of Emotional Context Insensitivity (ECI) [31, 32]. While the concept of ECI suggests a general dulling of emotional responses in depression, our mathematical definitions and analysis offer a more nuanced explanation. Our findings suggest that rather than simply a reduction in emotional intensity, which is also part of most other theories of affect disorders [7], depression is characterized by a blending of emotional boundaries, which accounts for the different theories of depressive emotions. This new perspective, grounded in dynamic systems theory, offers a more comprehensive understanding of emotional dynamics in depression, challenging and refining the traditional narratives often discussed in the literature.

Our study represents a pioneering approach in methodologically quantifying the dynamics of emotions over time through a principled analytical framework. A robust quantitative framework should be comprehensive, capable of explaining all aspects of the data without resorting to heuristics. While there is a growing interest in capturing the temporal dynamics of emotions [1, 2, 33, 34], existing measures often fall short in a critical area: they are often redundant, and their added value, particularly

in relation to psychological well-being remains unclear [6]. State space models on the other hand, as employed in our study, offer a universal framework that provides a non-redundant perspective on emotional dynamics [35]. Assuming linear interactions among emotions (with a discussion on non-linear extensions in our limitations section), the equations we have formulated represent the definitive model of state dynamics, as long as our data is enough. In this context, the time constants and controllability metrics we derived, alongside the equations of emotion evolution, are comprehensive tools for understanding emotional dynamics [11, 14]. Related, while measures like the temporal variance of emotional dynamics, a topic often explored in the literature, can be derived from our findings, the inverse relationship does not always hold. Consider, for example, the concept of emotional inertia, which describes how resistant emotional states are to change. This concept is typically assessed by examining the variance in emotional time series data. However, it is crucial to note that analyzing the variance of a single emotion in isolation might not provide a complete picture. As our research, along with other studies, has shown[36], emotions do not evolve in isolation but rather in conjunction with one another, making their covariance indicative of complex, multidimensional data characteristics. In contrast, the time constants we calculate offer a direct and precise insight into the unique characteristics of each emotion. This methodology allows for a more nuanced and distinct understanding of individual emotions, thereby enriching our grasp of the dynamic emotional landscape.

Recent behavioral studies, including one that examined responses to 2,185 videos [3, 36], reveal the complexity and breadth of our emotional world. This particular study highlighted that individuals can experience over 27 distinct emotions, which occupy a multi-faceted, high-dimensional space. To delve deep into the dynamics of these emotions, it is essential to monitor affective states over time. Tracking these emotional changes requires methodologies that can capture the fluctuations of feelings in everyday life. Among the most effective methods for this purpose are the Experience Sampling Method (ESM) and daily diary studies [37]. These approaches involve participants recording their emotional experiences at multiple points throughout their day-to-day lives, offering a window into the ever-changing landscape of human emotions. This ongoing monitoring provides valuable insights into how emotions evolve and interact over time, shedding light on the intricate nature of our emotional experiences. In these approaches, participants use mobile devices (like smartphones) to complete periodic questionnaires about their emotional experiences throughout the day. These methods offer a naturalistic assessment of emotional trajectories and reduce memory bias due to their proximity to real-time experiences. However, implementing these methods practically has its challenges [38]. Not all participants are comfortable using apps that require frequent updates throughout the day. Consequently, gathering sufficient data for meaningful analysis across large sample sizes has been increasingly difficult. For example, most recent studies have been limited to fewer than 100 participants and only span a few days [39]. To address this limitation and significantly increase our sample size, we turned to data from X (formerly Twitter), where people often voluntarily share their thoughts and emotions. Social media platforms like Twitter provide a semi-anonymous space where individuals frequently discuss their

mental health struggles and diagnoses. These platforms serve as a safe haven for sharing experiences, seeking support, and raising awareness about mental health issues [40, 41].

Prior online research has focused primarily on the negative emotions expressed by individuals with depression [42, 43]. However, it is important to note that the emotional life of someone with depression is not exclusively negative. For instance, in completing the Self-Rating of Happiness 2 scale [44], over 60% of individuals with depressive disorders reported feeling "completely happy", "very happy", or "quite happy", compared to over 90% in control groups [45]. Another study found that 68.4% of individuals with mental disorders often felt happy, versus 89.01% in a control group [46]. Thus, the aim should be to understand the temporal dynamics of emotions, particularly focusing on how they are influenced by positive life events.

In discussing the limitations and future directions of our study, it is important to acknowledge the constraints imposed by our current methodological approach and the opportunities for further research. One significant limitation of our study is the reliance on linear assumptions, which resulted in the characterization of emotional dynamics with a single time constant. However, it is plausible that emotions follow multiple time constants [47, 48], suggesting a more complex temporal structure. Future research could explore this by utilizing larger datasets, such as those available from Reddit [49], spanning several years. Employing non-linear analytical methods, such as Sparse Identification of Nonlinear Dynamics (SINDy) [50], could provide a more nuanced understanding of these multi-faceted emotional dynamics. Additionally, our study ventured into the realm of emotion dimensionality with an analysis of 28 different emotions, one of the widest ranges used in emotion studies to date. Despite this breadth, future studies could employ latent space models like Recurrent Neural Networks (RNNs) to explore emotional dimensions on a significantly larger scale [35, 51]. This approach could potentially increase the granularity of our emotional understanding by orders of magnitude. Related, our study focused exclusively on how emotional dynamics change in response to positive life events. However, it is crucial to extend this analysis to encompass a broader range of life events, including negative experiences. We hypothesize that the observed reduction in emotion controllability may not be directly linked to the valence of the experience, but this theory requires further investigation in future studies. Moreover, incorporating neuroimaging techniques could offer invaluable insights into the neural correlates of these emotional states, adding a biological perspective to our comprehension of emotional dynamics. Regarding the extension of our research, the application of NLP techniques could be instrumental [52, 53]. NLP could enable the extraction and analysis of various life events from large text corpora, thereby enhancing the scalability and depth of our study. Such an approach could provide a richer context for understanding how different life events influence emotional trajectories. Finally, a critical area for future investigation is the study of the impact of interventions, such as psychotherapy, on emotional dynamics. This line of inquiry is crucial for establishing causal relationships and understanding the effectiveness of various therapeutic approaches in modulating emotional states [51]. By examining the before-and-after effects of interventions on the emotional landscape,

particularly in clinical settings, we can move closer to developing targeted, evidence-based therapeutic strategies that are responsive to the unique emotional profiles of individuals.

4 Methods

4.1 Sample

We analyze the emotional ripple dynamics of the users from the Twitter dataset proposed by Shen et al. [54], further referred to in this work as Twitter-Depression. It is a large-scale dataset, and we used a balanced version of it [19] with tweets from 1,402 users with a depression diagnosis and 1,402 control users. The users from the depression group are individuals who mentioned in one of their social media posts that they were diagnosed with depression (e.g., "I was diagnosed with depression"). Control users are individuals without any mention of a depression diagnosis or of the word "depress" in their timeline. For constructing the dataset, posts within one month from the diagnosis mention were retrieved for both groups. In total, there are 292,564 posts in the depression group and 879,025 posts in the control group.

In the current work, two additional datasets were used: Twitter-STMHD [55] and HappyDB [26]. The purpose of the Twitter-STMHD dataset is to be used alongside HappyDB to train a model capable of detecting posts expressing happy moments. The Twitter-Depression dataset is kept intact for the final analyses. Twitter-STMHD [55] contains posts from individuals with eight mental disorders (e.g., depression, anxiety, PTSD, etc.), labeled by their mentions of diagnosis, similar to Twitter-Depression. However, given that Twitter-STMHD contains data from over 30,000 users with different mental disorders, for the purpose of our work, only a sample of the posts from the depression and control groups is used, totaling 215,000 posts.

The HappyDB database [26] is an extensive collection of 100,000 happy moments gathered from more than 10,000 crowdsourcing workers. Given that happy moments reported by individuals are not always portrayed using positive words [28], and cannot be reliably detected using emotion detection models [27], the HappyDB corpus is a valuable resource for understanding the expressions of happiness in natural language. The corpus contains individuals' responses to the question "What made you happy in the past 24 hours", or in the last three months. The collected happy moments come from a wide variety of categories, such as leisure (e.g., "We booked our beach vacation for May of this year.."), achievement (e.g., "Got A in class."), nature (e.g., "I went for a run in the neighborhood. I enjoyed the perfect weather."), exercising (e.g., "I had a great workout last night."). family (e.g., "I went to lunch with my girlfriend and her family."), friends (e.g., "I went out for dinner with my friends and enjoyed a lot."), and others. We use the cleaned version of the corpus, in which spelling mistakes are corrected. The texts underwent additional preprocessing, and a total of 90,641 happy moments from HappyDB were used in this work.

Data Preprocessing. The social media posts from Twitter-Depression and Twitter-STMHD were preprocessed by removing URLs, mentions, hashtags, and retweets. Non-English posts detected by the polyglot library [56] were removed. We further

filtered out the texts with less than 5 or more than 50 words, ensuring that the sentences from Twitter-Depression, Twitter-STMHD, and HappyDB are similar in terms of length.

4.2 Extracting Happy Moments and Emotions

To study the interaction of positive life events with other emotions expressed by individuals, we first identify the expressions of happy moments from the Twitter-Depression dataset corresponding to individuals with depression and control using the positive-unlabeled learning framework [57]. Next, we extract emotion information from each of the social media posts using a state-of-the-art transformer-based model [20].

Learning to Extract Happy Moments with Positive Unlabeled Learning. While highly performant methods for sentiment and emotion detection from natural language have been developed [20, 58, 59], finding the expressions of happy moments in social media remains a challenging task [27, 28]. We consider a happy moment to be a subjective experience characterized by feelings of joy, contentment, and satisfaction. For conveying happy moments in textual data, individuals might rely on language with neutral or even negative sentiment [28]. Given that popular methods for emotion detection do not generalize for happy moments identification [27], and there is no dataset of positive and negative data for training a binary classifier for the task, we develop a deep learning classifier following the positive-unlabeled (PU) learning framework [57, 60]. The PU learning framework offers a way of estimating a positive-negative classifier from positive and unlabeled data. PU learning aims to identify examples similar to a group of positive labeled examples in a large unlabeled dataset. Such an approach has been used in the past in areas such as disease gene identification [61, 62] and opinion mining [63]. In our work, we construct a PU dataset by combining samples from HappyDB and Twitter-STMHD. We use the instances with happy moments reports from HappyDB as the positive class and samples from Twitter-STMHD represent the unlabeled set, as they contain both positive and negative samples that are unknown at the training stage. We opted not to use Twitter-Depression in training, and only use it in the analysis of emotional dynamics.

Our goal is to extract happy moments statements from Twitter-Depression similar to those found in HappyDB. Methods for PU learning are based on different assumptions in their formulation, about the distribution of positives in the unlabeled set [60]. Notably, the Selected Completely at Random (SCAR) assumption [57] states that the labeled positive samples are representative of all the positive instances in the data, both labeled and unlabeled [64]. We impose the SCAR assumption in our work in the sense that we want the happy moments statements from Twitter-STMHD to be similar to the ones in HappyDB. In our setting, both the training and validation datasets contain positive and unlabeled data, and traditional performance metrics (i.e., Precision, Recall, F_1) cannot be correctly computed without knowing the true positive and negative labels. However, using the SCAR assumption allows us to evaluate the performance of the model without needing a manually annotated validation set with positive and negative labels [64]. Rather than computing performance metrics on the predicted

probability of an example being in the positive class, metrics are calculated on the predicted probability of an example being labeled. To choose the best-performing model during training, we estimate the F_1 score on the validation data under the SCAR assumption.

Usually, the PU-learning framework has two steps: estimating the fraction of positives from all the unlabeled samples (noted in this work with α) and incorporating the estimation into a positive-negative classifier. There have been other works that proposed methods to improve the performance of both steps [57, 65], but we opted to use a recent approach described by Garg et al. [66] called Transfer, Estimate and Discard $(TED)^n$. $(TED)^n$ offers both formal guarantees and good empirical results for PU Learning compared to other approaches. The algorithm iteratively uses Best Bin Estimation (BBE) and Conditional Value Ignoring Risk (CVIR) to improve the estimation α of positive samples in the unlabeled set and use this value to learn a positive-unlabeled classifier.

Formally, we aim to train a classifier $f_\theta : X \to [0, 1]$ on a PU Learning dataset $X = X_u \cup X_p$, with X_u representing the unlabeled set (in our case, Twitter-STMHD) and X_p representing the labeled set (in our case, HappyDB) by minimizing the expected cross-entropy loss l across the dataset. Given an estimate α through the BBE algorithm (see [66]), we set the gradient of each batch of examples i from both the labeled and unlabeled sets (X_p^i, X_n^i) as $\nabla_\theta[\alpha \cdot \hat{L}^+(f_\theta; X_p^i) + (1 - \alpha) \cdot \hat{L}^-(f_\theta); X_n^i]$, where $\hat{L}^+(f; X) = \sum_{i=1}^n l(f(x_i), +1)/n$ is the loss when predicting the samples as positive and $\hat{L}^-(f; X) = \sum_{i=1}^n l(f(x_i), -1)/n$ is the loss when predicting the samples as negative. The algorithm iteratively computes α, re-ranks samples according to their loss and updates the model with the gradients. We further provide details on the model architecture used, experimental setup and extraction of happy moments.

Transformer Architecture. In all our experiments for extracting happy moments and annotating them with emotions, we used variants of the transformer encoder model [67] pretrained on specific datasets [68, 69]. The transformer is a neural network architecture that has become the foundational model for natural language processing tasks [70, 71]. It was designed to work with sequential data by capturing pairwise dependencies between input elements and has proven to be substantially more performant than other network variants such as LSTMs [72] or GRUs [73]. Its success is due to a combination of design elements such as the self-attention mechanism (Eq. 1), multi-head attention (Eq. 2), residual connections [74] and layer normalization [75].

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{1}$$

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_h)W^O \tag{2}$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{3}$$

Furthermore, an important advantage compared to other types of networks is the ease with which the model can be computationally scaled up in terms of the number

of parameters through the parallelization of the internal multi-head attention operation. Transformers pretrained in a self-supervised manner on diverse and general datasets using, for example masked language modeling [70], offer substantial benefits through further fine-tuning on specific downstream problems [76, 77].

Experimental Setup. For the classification task of happy moments, we utilize a pretrained DistilBERT [68], which is a distilled version of BERT [70]. DistilBERT was chosen for its flexibility, as it is a smaller and faster transformer model that can be used for general-purpose tasks while still maintaining similar performance as the larger BERT model for downstream tasks. To train and validate our approach, we split the data into training and validation sets using a 9:1 ratio for both Twitter-STMHD and HappyDB. We trained the model for a maximum of 15 epochs, using a batch size of 200, a learning rate of 0.00005, and an initial $\alpha = 0.5$, similar to the setup of Garg et al. [66].

Model Selection. The model with the best performance is obtained by training the DistilBERT model for 6 epochs, with an estimated F_1 score of 98.59%. The high estimated performance of the model under the SCAR assumption shows that it can accurately predict the labeled positive examples from the data and keep the unlabeled positive instances to a minimum. The estimated proportion of happy moments in the unlabeled set was estimated to be $\alpha = 0.0013$. We further used the best-performing model to detect instances of happy moments from Twitter-Depression.

Extracting Happy Moments. We used the happy moments detection model on the texts from Twitter-Depression to classify if a tweet represents a happy moment or not. Each post of the user was labeled with a binary decision whether or not it contained a report of a happy moment. Out of the 1,402 users in the depression group, only 871 of them had at least one happy moment, while for the control group, only 819 of the 1,402 had at least one report of a happy moment. The individuals had an average of 5.7 happy moments. For further experiments, we only use data from the 1,690 users who reported at least one happy moment.

Emotion Estimation. We estimate emotions for all the social media posts in the span of one month for the 1,690 users from the Twitter-Depression dataset. To this end, we use a state-of-the-art transformer-based EmoRoBERTa model [20]. EmoRoBERTa is a RoBERTa [78] model trained on the GoEmotions dataset [69] to predict 28 emotions, including neutral. The EmoRoBERTa model is used to estimate the probability distribution across the 28 emotions for each social media post. Consequently, each post is automatically annotated with probability scores for each emotion.

4.3 Dynamic System Theory

Following previous work [14], we assume the psychological behavior to follow a noise-free linear time-invariant model given by

$$x(k+1) = Ax(k) + Bu(k) \tag{4}$$

where $x(k) \in D^{28}$, $D \in 0, 1$ are the state variables and here correspond to the emotional readings at time k, and $u(k) \in [0, 1]$ is the input variable and here corresponds to the existence of a positive life event at time k. In this formulation, the evolution of the emotion variables (i.e., $x(k)$) is determined by a weighted summation of the emotional values and the input variables. Concrete, A represents the interaction between emotions and B specifies how the intervention is the evolution of emotions. Importantly, following previous work [14], we assume that A remains constant. In these models, we estimated $A_{28\times28}$ and $B_{28\times1}$ based on Dynamic Mode Decomposition with Control (DMDc[15]). In the most straightforward implementation which we used in this paper, defining

$$X_1 = \begin{bmatrix} x_1 & x_2 & \dots & x_{m-1} \end{bmatrix} \tag{5}$$

$$X_2 = \begin{bmatrix} x_2 & x_3 & \dots & x_m \end{bmatrix} \tag{6}$$

$$U = \begin{bmatrix} u_1 & u_2 & \dots & u_{m-1} \end{bmatrix} \tag{7}$$

where $x_k = x(k)$ and $u_k = u(k)$, we can rewrite equation 4 and thus solve for A and B simultaneously as follows:

$$X_2 = \begin{bmatrix} A & B \end{bmatrix} \begin{bmatrix} X_1 \\ U \end{bmatrix} \tag{8}$$

$$\begin{bmatrix} A & B \end{bmatrix} = X_2 \begin{bmatrix} X_1 \\ U \end{bmatrix}^{\dagger} \tag{9}$$

Where † denotes Moore–Penrose pseudoinverse [79]. Based on these models, we can compute the following metrics:

Emotion Reactivity. We operationalized Emotion Reactivity through Average Controllability (AC) which is a metric used to quantify the ease with which a node's state can be altered through the use of input controls, providing insights into the influence of individual nodes within the network. It is mathematically defined as:

$$AC_j = \text{trace}\left(\sum_{i=0}^{\infty} A^i B_j B_j^T (A^T)^i\right) \tag{10}$$

where A is defined as before and B_j is the j^{th} canonical vector. This measure was instrumental in assessing the influence of individual nodes within the network. Average controllability quantifies how easily the state of a node (in this case, an emotion) can be changed through external inputs. It is a measure of the potential influence that can be exerted over a node to drive the system to different states. In the context of our

study, it represents how easily an emotion can be influenced or controlled. Conceptually, average controllability is associated with the averaged interconnections between nodes, wherein nodes exhibiting higher average controllability are those for which interventions result in more pronounced changes around their current values [80].

Time Constant. The time constant, typically denoted as τ, measures the system's speed of response and is crucial for understanding the transient dynamics of the system. It is defined as the inverse of the system's eigenvalues, $\tau = \frac{1}{\lambda}$, and it provides insights into how quickly the system reacts to changes in input and initial conditions. The Time Constant (τ) of a system is a measure of the time required for the system to reach a specific state or to undergo a particular process. In the context of emotional dynamics, it can be interpreted as the amount of time required for an emotion to reach a stable state or to undergo a significant change.

Control energy. In linear dynamical systems, the Gramian matrix helps estimate the energy needed for state changes. The controllability Gramian, W_c, is given by:

$$W_c = \sum_{i=0}^{\infty} A^i B_j B_j^T (A^T)^i \tag{11}$$

where A is the state matrix and B is the input matrix. This matrix reveals how inputs affect the system's state.

As detailed in [30], the minimum and average eigenvalues of the controllability Gramian play a pivotal role. They measure the amount of control energy necessary to reach a specified state. Networks characterized by smaller Gramian eigenvalues demand a greater amount of energy for control, a significant factor when energy resources are constrained.

5 Availability of data and codes

The codes to replicate the simulations will become publicly available upon acceptance of this manuscript at https://github.com/PsyControLab. The data used in this study is publicly available at https://zenodo.org/records/6409736 (Twitter-STMHD), https://drive.google.com/file/d/11ye00sHFY5re2NOBRKreg-tVbDNrc7Xd/ (Twitter-Depression).

6 Funding

HJ was supported by von Behring-Röntgen Stiftung (No. 70-00038). PR was supported by MCIN/AEI/10.13039/501100011033 and by ERDF, EU A way of making Europe (No. PID2021-124361OB-C31). The work of LF has been supported by the German Federal Ministry of Education and Research (BMBF) as a part of the Junior AI Scientists program under the reference 01-S20060.

7 Conflict of Interest

The Authors declare no competing interests.

8 Author contribution statement

Conceptualization: HJ, AB, LF, PR Methodology and Validation: AB, AC, HJ, TK, Writing-Original Draft: AB, AC, HJ, TK, Data: AB, AC, PR, LF, Writing-Review: all.

References

[1] Gross, J.J.: Emotion regulation: Current status and future prospects. Psychological inquiry **26**(1), 1–26 (2015)

[2] Kuppens, P., Verduyn, P.: Emotion dynamics. Current Opinion in Psychology **17**, 22–26 (2017)

[3] Horikawa, T., Cowen, A.S., Keltner, D., Kamitani, Y.: The neural representation of visually evoked emotion is high-dimensional, categorical, and distributed across transmodal brain regions. Iscience **23**(5), 101060 (2020)

[4] Koval, P., Pe, M., Meers, K., Kuppens, P.: Affect dynamics in relation to depressive symptoms: Variable, unstable or inert? Journal of Abnormal Psychology **122**(3), 684–694 (2013)

[5] Oravecz, Z., Tuerlinckx, F., Vandekerckhove, J.: A hierarchical latent stochastic differential equation model for affective dynamics. Psychological Methods **16**(4), 468–490 (2011)

[6] Dejonckheere, E., Mestdagh, M., Houben, M., Rutten, I., Sels, L., Kuppens, P., Tuerlinckx, F.: Complex affect dynamics add limited information to the prediction of psychological well-being. Nature human behaviour **3**(5), 478–491 (2019)

[7] Bylsma, L.M., Morris, B.H., Rottenberg, J.: A meta-analysis of emotional reactivity in major depressive disorder. Clinical Psychology Review **28**(4), 676–691 (2008)

[8] Rottenberg, J., Gross, J.J., Gotlib, I.H.: Emotion context insensitivity in major depressive disorder. Journal of abnormal psychology **114**(4), 627 (2005)

[9] Christensen, M.C., Ren, H., Fagiolini, A.: Emotional blunting in patients with depression. part i: clinical characteristics. Annals of General Psychiatry **21**(1), 1–8 (2022)

[10] Bringmann, L.F., Pe, M.L., Vissers, N., Ceulemans, E., Borsboom, D., Vanpaemel, W., Tuerlinckx, F., Kuppens, P.: Assessing temporal emotion dynamics using networks. Assessment **23**(4), 425–435 (2016)

[11] Kirk, D.E.: Optimal control theory: an introduction. Courier Corporation (2004)

[12] Camras, L., Witherington, D.C.: Dynamical systems approaches to emotional development. Developmental Review **25**(3-4), 328–350 (2005)

[13] Jamalabadi, H., Hofmann, S.G., Teutenberg, L., Emden, D., Goltermann, J., Enneking, V., Meinert, S., Koch, K., Leehr, E., Stein, F., et al.: A complex systems model of temporal fluctuations in depressive symptomatology (2022)

[14] Stocker, J.E., Koppe, G., Reich, H., Heshmati, S., Kittel-Schneider, S., Hofmann, S.G., Hahn, T., Maas, H.L., Waldorp, L., Jamalabadi, H.: Formalizing psychological interventions through network control theory. Scientific Reports **13**(1), 13830 (2023)

[15] Proctor, J.L., Brunton, S.L., Kutz, J.N.: Dynamic mode decomposition with control. SIAM Journal on Applied Dynamical Systems **15**(1), 142–161 (2016)

[16] Koval, P., Brose, A., Pe, M.L., Houben, M., Erbas, Y., Champagne, D., Kuppens, P.: Emotional inertia and external events: The roles of exposure, reactivity, and recovery. Emotion **15**(5), 625 (2015)

[17] Kuppens, P., Allen, N.B., Sheeber, L.B.: Emotional inertia and psychological maladjustment. Psychological science **21**(7), 984–991 (2010)

[18] Kuppens, P., Sheeber, L.B., Yap, M.B., Whittle, S., Simmons, J.G., Allen, N.B.: Emotional inertia prospectively predicts the onset of depressive disorder in adolescence. Emotion **12**(2), 283 (2012)

[19] An, M., Wang, J., Li, S., Zhou, G.: Multimodal topic-enriched auxiliary learning for depression detection. In: Proceedings of the 28th International Conference on Computational Linguistics, pp. 1078–1089 (2020)

[20] Kamath, R., Ghoshal, A., Eswaran, S., Honnavalli, P.: An enhanced context-based emotion detection model using roberta. In: 2022 IEEE CONECCT, pp. 1–6 (2022). IEEE

[21] Bucur, A.-M., Cosma, A., Dinu, L.P.: Life is not always depressing: Exploring the happy moments of people diagnosed with depression. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference, pp. 4186–4192 (2022)

[22] Bucur, A.-M., Chulvi, B., Cosma, A., Rosso, P.: The expression of happiness in social media of individuals diagnosed with depression. Authorea Preprints (2023) https://doi.org/10.36227/techrxiv.23264507.v1

[23] Beck, A.T., Steer, R.A., Brown, G.K.: Bdi-ii: Beck depression inventory. Pearson (1996)

[24] Tang, E., Bassett, D.S.: Colloquium: Control of dynamics in brain networks. Reviews of modern physics **90**(3), 031003 (2018)

[25] Leppänen, J.M., Milders, M., Bell, J.S., Terriere, E., Hietanen, J.K.: Depression biases the recognition of emotionally neutral faces. Psychiatry research **128**(2), 123–133 (2004)

[26] Asai, A., Evensen, S., Golshan, B., Halevy, A., Li, V., Lopatenko, A., Stepanov, D., Suhara, Y., Tan, W.-C., Xu, Y.: Happydb: A corpus of 100,000 crowdsourced happy moments. In: Proceedings of Language Resources and Evaluation Conference (2018)

[27] Jaidka, K., Chhaya, N., Mumick, S., Killingsworth, M., Halevy, A., Ungar, L.: Beyond positive emotion: Deconstructing happy moments based on writing prompts. In: Proceedings of International AAAI Conference on Web and Social Media, vol. 14, pp. 294–302 (2020)

[28] Moreno-Ortiz, A., Pérez-Hernández, C., García-Gámez, M.: The language of happiness in self-reported descriptions of happy moments: Words, concepts, and entities. Humanities and social sciences communications **9**(1), 1–13 (2022)

[29] Henry, T.R., Robinaugh, D.J., Fried, E.I.: On the control of psychological networks. Psychometrika **87**(1), 188–213 (2022)

[30] Pasqualetti, F., Zampieri, S., Bullo, F.: Controllability metrics, limitations and algorithms for complex networks. IEEE Transactions on Control of Network Systems **1**(1), 40–52 (2014)

[31] Rottenberg, J., Hindash, A.C.: Emerging evidence for emotion context insensitivity in depression. Current Opinion in Psychology **4**, 1–5 (2015) https://doi.org/10.1016/j.copsyc.2014.12.025

[32] Bylsma, L.M.: Emotion context insensitivity in depression: Toward an integrated and contextualized approach. Psychophysiology **58**(2), 13715 (2021)

[33] Houben, M., Van Den Noortgate, W., Kuppens, P.: The relation between short-term emotion dynamics and psychological well-being: A meta-analysis. Psychological bulletin **141**(4), 901 (2015)

[34] Goicoechea, C., Dakos, V., Sanabria, D., Heshmati, S., Westhoff, M., Banos, O., Pomares, H., Hofmann, S.G., Perakakis, P.: Shifts in affect dynamics predict psychological well-being (2023)

[35] Durstewitz, D., Koppe, G., Thurm, M.I.: Reconstructing computational system dynamics from neural data with recurrent neural networks. Nature Reviews Neuroscience, 1–18 (2023)

[36] Cowen, A.S., Keltner, D.: Self-report captures 27 distinct categories of emotion bridged by continuous gradients. Proceedings of the national academy of sciences **114**(38), 7900–7909 (2017)

[37] Trull, T.J., Ebner-Priemer, U.W.: Using experience sampling methods/ecological momentary assessment (esm/ema) in clinical assessment and clinical research: introduction to the special section. (2009)

[38] Smith, K.A., Blease, C., Faurholt-Jepsen, M., Firth, J., Van Daele, T., Moreno, C., Carlbring, P., Ebner-Priemer, U.W., Koutsouleris, N., Riper, H., et al.: Digital mental health: challenges and next steps. BMJ Ment Health **26**(1) (2023)

[39] Haslbeck, J., Ryan, O., Dablander, F.: Multimodality and skewness in emotion time series. Emotion (2023)

[40] Berry, N., Lobban, F., Belousov, M., Emsley, R., Nenadic, G., Bucci, S.: # whywetweetmh: understanding why people use twitter to discuss mental health problems. Journal of medical Internet research **19**(4), 107 (2017)

[41] De Choudhury, M., De, S.: Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 8, pp. 71–80 (2014)

[42] Park, M., Cha, C., Cha, M.: Depressive moods of users portrayed in twitter. In: Proceedings of the 18th ACM International Conference on Knowledge Discovery and Data Mining, pp. 1–8 (2012)

[43] Chen, X., Sykora, M.D., Jackson, T.W., Elayan, S.: What about mood swings: Identifying depression on twitter with temporal measures of emotions. In: Companion Proceedings of the Web Conference 2018, pp. 1653–1660 (2018)

[44] Veenhoven, R., Ehrhardt, J.: The cross-national pattern of happiness: Test of predictions implied in three theories of happiness. Social indicators research **34**, 33–68 (1995)

[45] Spinhoven, P., Elzinga, B.M., Giltay, E., Penninx, B.W.: Anxious or depressed and still happy? PloS one **10**(10), 0139912 (2015)

[46] Bergsma, A., Have, M.t., Veenhoven, R., Graaf, R.d.: Most people with mental disorders are happy: A 3-year follow-up in the dutch general population. The Journal of Positive Psychology **6**(4), 253–259 (2011)

[47] Heshmati, S., Sbarra, D.A., Benson, L.: Integrating multiple time-scales to advance relationship science. Journal of Social and Personal Relationships **40**(4), 1069–1078 (2023)

[48] Heshmati, S., Muth, C., Roeser, R.W., Smyth, J., Jamalabadi, H., Oravecz,

Z.: Conceptualizing psychological well-being as a dynamic process: Implications for research on mobile health interventions. Social and Personality Psychology Compass **18**(1), 12933 (2024)

[49] Yates, A., Cohan, A., Goharian, N.: Depression and self-harm risk assessment in online forums. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2968–2978 (2017)

[50] Brunton, S.L., Proctor, J.L., Kutz, J.N.: Discovering governing equations from data by sparse identification of nonlinear dynamical systems. Proceedings of the national academy of sciences **113**(15), 3932–3937 (2016)

[51] Jamalabadi, H., Koppe, G., Nozari, E., Hahn, T.: Engineering the mind: Actionable technical requirements for innovation in clinical neuroscience (2023)

[52] Schneider, K., Leinweber, K., Jamalabadi, H., Teutenberg, L., Brosch, K., Pfarr, J.-K., Thomas-Odenthal, F., Usemann, P., Wroblewski, A., Straube, B., *et al.*: Syntactic complexity and diversity of spontaneous speech production in schizophrenia spectrum and major depressive disorders. Schizophrenia **9**(1), 35 (2023)

[53] Stewart, R., Velupillai, S.: Applied natural language processing in mental health big data. Neuropsychopharmacology **46**(1), 252 (2021)

[54] Shen, G., Jia, J., Nie, L., Feng, F., Zhang, C., Hu, T., Chua, T.-S., Zhu, W., *et al.*: Depression detection via harvesting social media: A multimodal dictionary learning solution. In: IJCAI, pp. 3838–3844 (2017)

[55] Singh, A.K., Arora, U., Shrivastava, S., Singh, A., Shah, R.R., Kumaraguru, P., *et al.*: Twitter-stmhd: An extensive user-level database of multiple mental health disorders. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 16, pp. 1182–1191 (2022)

[56] Al-Rfou', R., Perozzi, B., Skiena, S.: Polyglot: Distributed word representations for multilingual NLP. In: Proceedings of CoNLL, pp. 183–192 (2013)

[57] Elkan, C., Noto, K.: Learning classifiers from only positive and unlabeled data. In: Proceedings of ACM SIGKDD, pp. 213–220 (2008)

[58] Wankhade, M., Rao, A.C.S., Kulkarni, C.: A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review **55**(7), 5731–5780 (2022)

[59] Sailunaz, K., Dhaliwal, M., Rokne, J., Alhajj, R.: Emotion detection from text and speech: a survey. Social Network Analysis and Mining **8**, 1–26 (2018)

[60] Bekker, J., Davis, J.: Learning from positive and unlabeled data: a survey.

Machine learning **109**(4), 719–760 (2020)

[61] Mordelet, F., Vert, J.-P.: Prodige: Prioritization of disease genes with multitask machine learning from positive and unlabeled examples. BMC bioinformatics **12**, 1–15 (2011)

[62] Mordelet, F., Vert, J.-P.: A bagging svm to learn from positive and unlabeled examples. Pattern recognition letters **37**, 201–209 (2014)

[63] Hernández-Fusilier, D., Montes-y-Gómez, M., Rosso, P., Guzmán-Cabrera, R.: Detecting positive and negative deceptive opinions using pu-learning. Information processing & management **51**(4), 433–443 (2015)

[64] Saunders, J.D., Freitas, A.A.: Evaluating the predictive performance of positive-unlabelled classifiers: a brief critical review and practical recommendations for improvement. ACM SIGKDD Explorations Newsletter **24**(2), 5–11 (2022)

[65] Du Plessis, M.C., Niu, G., Sugiyama, M.: Analysis of learning from positive and unlabeled data. Advances in neural information processing systems **27** (2014)

[66] Garg, S., Wu, Y., Smola, A.J., Balakrishnan, S., Lipton, Z.: Mixture proportion estimation and pu learning: a modern approach. Advances in Neural Information Processing Systems **34**, 8532–8544 (2021)

[67] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[68] Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019)

[69] Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., Ravi, S.: Goemotions: A dataset of fine-grained emotions. In: Proceedings of Annual Meeting of the Association for Computational Linguistics, pp. 4040–4054 (2020)

[70] Kenton, J.D.M.-W.C., Toutanova, L.K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 4171–4186 (2019)

[71] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., *et al.*: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45 (2020)

[72] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation

9(8), 1735–1780 (1997)

[73] Cho, K., Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, pp. 1724–1734 (2014)

[74] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

[75] Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., Zhang, H., Lan, Y., Wang, L., Liu, T.: On layer normalization in the transformer architecture. In: Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 10524–10533 (2020)

[76] Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Las Casas, D., Hendricks, L.A., Welbl, J., Clark, A., *et al.*: An empirical analysis of compute-optimal large language model training. Advances in Neural Information Processing Systems **35**, 30016–30030 (2022)

[77] Hernandez, D., Kaplan, J., Henighan, T., McCandlish, S.: Scaling laws for transfer. arXiv preprint arXiv:2102.01293 (2021)

[78] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)

[79] Barata, J.C.A., Hussein, M.S.: The moore–penrose pseudoinverse: A tutorial review of the theory. Brazilian Journal of Physics **42**, 146–165 (2012)

[80] Karrer, T.M., Kim, J.Z., Stiso, J., Kahn, A.E., Pasqualetti, F., Habel, U., Bassett, D.S.: A practical guide to methodological considerations in the controllability of structural brain networks. Journal of neural engineering **17**(2), 026031 (2020)