Compare Results

Old File:		New File:
2024.emnlp-main.1069.pdf	versus	**2024_emnlp-main_1069.pdf**
18 pages (298 KB)		**15 pages (344 KB)**
10/31/2024 10:33:09 PM		2/9/2026 12:44:34 PM

Total Changes

1197

Content

119 Replacements

106 Insertions

245 Deletions

Styling and Annotations

426 Styling

301 Annotations

Go to First Change (page 2)

Emotion Granularity from Text: A Novel Indicator for Mental Health Conditions

Anonymous Authors
1Department of Computer Science, University of Toronto
2Vector Institute, Toronto
3Department of Computing Science, University of Alberta
4Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill
5National Research Council Canada

Abstract

We are united in how emotions are central to shaping our experiences; yet, individuals differ greatly in how we each identify, categorize, and express emotions. In psychology, variation in the ability of individuals to differentiate between emotion concepts is called emotion granularity (determined through self- reports of one's emotions). High emotion granularity has been linked with better mental and physical health; whereas low emotion granularity has been linked with maladaptive emotion regulation strategies and poor health outcomes. In this work, we propose computational measures of emotion granularity derived from temporally- ordered speaker utterances in social media (in lieu of self- reports that suffer from various biases). We then investigate the effectiveness of such text- derived measures of emotion granularity in functioning as markers of various mental health conditions (MHCs). We establish baseline measures of emotion granularity derived from textual utterances, and show that, at an aggregate level, emotion granularities are significantly lower for people self- reporting as having an MHC than for the control population. This paves the way towards a better understanding of the MHCs, and specifically the role emotions play in our well-being.

1 Introduction

text[[115, 728, 487, 918], [511, 260, 881, 389]] Emotions play a central role in how we construct meaning and communicate with those around us. Yet, individuals vary in their understanding and experience of emotions, or "emotional expertise" (Hoemann et al., 2021b). Some people are able to recognize, identify, and describe what they feel using precise, context-specific terms like guilt, anger, frustration, or helplessness; others tend to use more broad terms to convey a general sense of feeling bad or feeling low. Emotion granularity (EG), aka emotion differentiation, is defined by psychologists as the ability of an individual to experience and categorize emotions in very specific terms (Barrett et al., 2001). Highly granular individuals have a broad range of highly situated and differentiated emotion concepts, and can reliably describe these concepts using language — for example, distinguishing between when they are feeling angry vs. when they are feeling sad, or when they are feeling elated from when they are feeling content.

Evidence collected in the last two decades provides consistent support for a link between emotional granularity and mental health (Erbas et al., 2014, 2018; Starr et al., 2017; Seah et al., 2020), physical health (Hoemann et al., 2021b; Bonar et al., 2023), and adaptive health behavior (Dixon- Gordon et al., 2014; Kashdan et al., 2015). Note that this is different from other findings that study how the prevalence

of specific emotions varies with mental health, (for example, people with depression tend to use more sadness- associated words). The link between EG and mental health suggests that there is a fundamental difference in how one perceives an emotion (broadly or specifically), and that in turn can impact their mental health.

Typically, granularity is measured across emotions with the same valence; one can therefore have a measure of negative emotion granularity, measured as the granularity of negative emotions (such as anger, sadness, and fear) and positive emotion granularity, measured as the granularity of positive emotions (such as joy, excitement, and satisfaction). Some works also look at the co- endorsement of emotions that express opposite valence, such as joy and sadness (Lindquist and Barrett, 2008).

In psychology and the affective sciences, emotion granularity is often measured using repeated measurements, where individuals are asked to rate the intensity of experiencing certain emotions multiple times over a period of days (e.g., 2- 3 times each day for 5 days), i.e, with self- reports of emotions felt. An individual's emotional granularity is then operationalized as the extent to which multiple emotions are co-endorsed over time, i.e, how similarly the emotions are rated across all measurements, using the intraclass correlation coefficient (ICC) (Shrout and Fleiss, 1979), which measures the extent to which the emotions covary in reports at the aggregate level. Individuals who tend to frequently rate multiple emotions at the same intensity levels are defined as low in granularity — the frequent co- endorsement across time indicates that they are failing to differentiate between these emotions in their reports. In contrast, individuals high in emotion granularity co- endorse multiple emotions less frequently over time (Tugade et al., 2004; Hoemann et al., 2021a; Lee et al., 2017; Reitsema et al., 2022).

While prior work in NLP has studied the link between emotions and mental health, these have largely been limited to measuring the prevalence or intensity of positive and negative emotions. In this work, we, for the first time, propose a way to compute emotion granularity from the textual utterances of an individual. Our method uses the temporal sequence of the utterances to first construct emotion arcs along multiple emotions, and computes granularity as the correlation of these emotion arcs. We hypothesize that this measure is indicative of the individual consistently expressing the same set of emotions together over a period of time, and can therefore act as a proxy measure of emotional granularity.

We then study the relationship between aggregate, population- level measures of emotion granularity in text for eight Mental Health Conditions (MHCs), namely attention- deficit hyperactivity disorder (ADHD), anxiety, bipolar disorder, depression, major depressive disorder (MDD), obsessive- compulsive disorder (OCD), postpartum depression (PPD), and post- traumatic stress disorder (PTSD), and compare them to a control group. We use two social media datasets where users have chosen to self- disclose their mental health diagnosis (Suhavi et al., 2022; Losada et al., 2017, 2018). We compute emotion granularity metrics for each of these groups to answer the following questions:

1. Do measures of emotional granularity differ between the MHCs and the control group?
2. Which measures of emotion granularity are the most effective at differentiating between the MHCs and the control group?
3. Which emotion pairs lead to the greatest difference in granularity between an MHC and the control group?

Exploring this line of questions helps us better understand how emotion granularity presents itself in text, whether emotion granularity from text can be a useful tool to study MHCs, and how an MHC impacts our perception of emotions (and perhaps even, how the perception of emotions impacts our mental health).

Our results establish baseline measures of emotion granularity from text, and show that these measures function as reliable indicators, at the aggregate- level, for the presence of many of the mental health conditions we study. Our work makes an important contribution to the growing wealth of research on textual measures of emotional expression as biosocial markers of MHCs, and has a broader utility in functioning as an additional indicator of the mental well- being of

populations.[1]

2 Related Work

2.1 Emotions and Mental Health

Measures of emotional experience and their patterns of change over time have been extensively studied as markers of mental and physical wellbeing (Lewis et al., 2010). The Emotion Dynamics framework in psychology quantifies the patterns with which emotions change over time, allowing researchers to better understand emotional experiences and individual variation (Kuppens and Verduyn, 2017). The framework includes several measures such as the duration, intensity, variability, and granularity of one's emotional experiences. Numerous studies in psychology have shown emotion dynamics correlate with overall wellbeing, mental health, and psychopathology (the scientific study of mental illness or disorders) (Kuppens and Verduyn, 2017; Houben et al., 2015; Silk et al., 2011; Sperry et al., 2020).

Emotion granularity in particular is positively associated with adaptive behaviour in adverse conditions — accurately labeling our emotions can inform us of the right coping strategies to use in different contexts. Individuals with higher emotion granularity tend to use a broader range of strategies to deal with negative emotions, and are more successful at doing so (Barrett et al., 2001). Several studies have shown that emotion granularity is lower in individuals with mental health conditions like bipolar disorder (Suvak et al., 2011; Dixon- Gordon et al., 2014), manic depressive disorder (Demiralp et al., 2012), schizophrenia (Kring et al., 2003), autism spectrum disorder (Erbas et al., 2013), and affective disorders like anxiety (Seah et al., 2020) and depression (Starr et al., 2017; Willroth et al., 2020). Lower granularity is also associated with increased tendencies to engage in maladaptive behaviour, such as alcohol consumption (Kashdan et al., 2015; Emery et al., 2014) and aggression (Pond Jr et al., 2012).

Researchers in affective science typically measure emotional granularity through experience sampling methodologies (ESMs), or ecological momentary assessments (EMAs), where individuals (participants) are repeatedly asked to report on their emotional states on several occasions throughout the day, for several days. For example, participants may be asked to endorse a series of ten emotion words (e.g., anger, fear, happy, etc.) on a Likert scale across several sampling instances. Emotional granularity would then be computed as the intraclass correlation (ICC) of ratings across sampling instances. A high ICC would suggest that a participant experiences all of the emotions in a similar way across trials (treating them as synonyms for more general affectual states such as "unpleasantness" or "pleasantness"), whereas a low ICC would suggest that a participant experienced emotions in a granular and context- specific way.

While emotion granularity is generally measured between emotion categories that are close to each other in the affective space (i.e., express similar valence), the concept of dialecticism refers to the co- incidental experience of both negative and positive emotions (Lindquist and Barrett, 2008). Dialecticism can therefore be operationalized as the co- endorsement of emotion pairs that express positive and negative valence.

2.2 Language and Mental Health

text[[115, 808, 486, 918], [511, 84, 881, 180]] Given the limitations of self- report surveys (e.g., limited data coverage and time spans, biases, etc (Kragel et al., 2022)), another approach to measure well- being indicators is through one's language usage. Some well-known linguistic indicators of mental health include the proportion of pronouns used for those with depression (Koops et al., 2023), syntax reduction for anorexia nervosa (Cuteri et al., 2022), certain lexical and syntactic features for mild cognitive impairment and dementia (Cal et al., 2021; Gagliardi and Tamburini, 2021), and semantic connectedness for schizophrenia (Corcoran et al., 2020).

Recently, another linguistic feature that researchers leveraged for insights into overall wellbeing, are the emotions expressed in language. Largely, only sentiment has been explored and mainly from social media

[1] All our code will be made available through the project webpage.

data (a rich source of language data). For example, more negative sentiment was expressed in text by individuals with depression (De Choudhury et al., 2013; Seabrook et al., 2018; De Choudhury et al., 2021). Other work has found that suicide watch, anxiety, and self- harm subreddits had markedly lower negative sentiment compared to other mental health subreddits such as Autism and Asperger's (Gkotsis et al., 2016).

Hipson and Mohammad (2021) and Vishnubhotla and Mohammad (2022) introduced Utterance Emotion Dynamics (UED), a framework to quantify patterns of change of emotional states associated with utterances along a longitudinal (temporal) axis (using data from screenplays and tweets). Teodorescu et al. (2023) found that measures of emotion dynamics from text correlate with various mental health diagnoses.

These works overall show that the average emotion expressed in text and also the characteristics of individual emotion change over time (e.g., variability) are meaningful indicators of well- being. In this work, we explore whether the degree of co- expression of pairs of emotions in text (emotion granularity) is a meaningful indicator of mental health.

3 Datasets

We use the Twitter- STMHD dataset (Suhavi et al., 2022) for our experiments and also verify our results with a smaller Reddit eRisk (Losada et al., 2017, 2018) dataset. We describe both of them below.

Twitter- STMHD dataset: Suhavi et al. (2022) identified tweeters who self- disclosed as having an MHC diagnosis using carefully constructed regular expression patterns and manual verification. We summarize key details on the dataset creation process in Appendix A. The control group consists of users identified from a random sample of tweets (who posted during approximately the same time period as the MHC tweets). These tweeters did not post any tweets meeting the MHC regex described above. Additionally, users who had any posts about mental health discourse were removed from the control group. Note that this process does not guarantee that users in the control group did not have an MHC diagnosis, but rather the group as a whole may have very few tweeters from these MHC groups. The number of users in the control group was selected to match the size of the depression dataset, which had the largest number of users.

For the final set of users, four years of tweets were collected for each user: two years before selfreporting a mental health diagnosis and two years after. For the control group, tweets were randomly sampled from between January 2017 and May 2021 (same date range as the other MHC classes).

Reddit eRisk dataset: To further add to our findings, we also included the eRisk 2018 dataset (Losada et al., 2017, 2018) in our experiments. It consists of users who self- disclosed as having depression on Reddit (expressions were manually checked), and a control group (individuals were randomly sampled). The dataset includes several hundred posts per user, over approximately a 500- day period. We combined users and their instances from both the training set (which is from the eRisk 2017 task (Losada et al., 2017)) and the test set (from the eRisk 2018 task (Losada et al., 2018)).

3.1 Preprocessing

We further preprocessed both the Twitter- STMHD dataset and the eRisk dataset for our experiments (Section 4), as we are specifically interested in the relationship between emotion granularity and each disorder. Several users self- reported as being diagnosed with more than one disorder, referred to as comorbidity. We found a high comorbidity rate between users who self- reported as having anxiety and depression, as is also supported in the literature (Pollack, 2005; Gorman, 1996; Hirschfeld, 2001; Cummings et al., 2014). Since we wanted to focus on each MHC separately (and not on co- morbidity) we only considered users who self- reported as having one MHC. We also performed the following preprocessing steps:

We only considered posts in English. We filtered out posts that contained URLs (the text in such posts is often not self- contained). We removed retweets (identified through tweets containing RT, rt). This is to focus exclusively on texts written by the user. To ensure that we did not include users that post very infrequently or very frequently, we excluded users

based on the number of posts per individual. We discarded data from those who either had less than 100 posts (as was similarly done in Vishnubhotla and Mohammad (2022)) and those who had posted more than 1.5 times the interquartile range above quartile three (75th percentile) of the control group.[2]

Table 1: The number of users in each MHC, the average number of posts per user, and the average number of tokens per post in the preprocessed version of the Twitter-STMHD and Reddit eRisk datasets.

Dataset, Group	#people	Av. #posts	Av. #tokens per post
Twitter			
MHC	19,324	2,590.48	17.59
ADHD	6,356	2,497.43	17.46
Anxiety	3,036	2,921.05	17.46
Bipolar	1,061	2,820.17	17.32
Depression	4,855	2,526.62	16.75
MDD	219	2,640.69	16.40
OCD	1,009	2,388.73	18.38
PPD	179	2,581.19	19.18
PTSD	2,609	2,533.85	19.41
Control	6,001	2,420.50	16.16
Reddit			
Depression	112	556.57	47.22
Control	90	7665.00	41.09

4 Emotion Granularity from Text

The core metric that we want to capture from the text utterances of an individual is emotion granularity-what psychologists term the "coendorsement" of pairs of emotions. Analogous to their operationalization of granularity in terms of the Intra- Class Correlation (ICC) of repeated emotion intensity measurements along emotion adjectives, we use textual utterances to derive a temporal sequence of emotion states, referred to as an emotion arc for the speaker (section 4.1), and operationalize granularity as the correlations of these arcs. We construct emotion arcs for multiple emotions, for each user in the MHC groups and the control group.

Emotion Dimensions: A key requirement of our computational method is that we must be able to quantify the emotional score of a text along a selected emotion dimension. We are therefore limited by the resources and models available to compute such a score for an emotion dimension. Here, keeping in mind the necessity of including multiple emotion dimensions that are similarly-valenced, we work with the eight emotions represented in the NRC Emotion Intensity Lexicon: anger, anticipation, disgust, fear, joy, sadness, surprise, and trust (Mohammad, 2018). We partition these emotions into three groups based on valence association: joy and trust are in the positive valence group; anger, sadness, fear, and disgust are in the negative valence group; and anticipation and surprise are in the variable valence group. The distinction for surprise and anticipation is necessary because specific instances of these emotions can have a positive or a negative connotation (e.g., a good or a bad surprise).

4.1 Constructing Emotion Arcs

We order the utterances for each user based on timestamp information in the metadata. We construct emotion arcs for the temporal sequence of utterances of each user, along each of the eight emotions, in two ways pertaining to different window sizes. This is to make sure that the results are largely robust even when varying the window size to some extent.

Utterance-level Window: Emotion scores (for each emotion category) are computed for each utterance (i.e, tweet or Reddit post). Here, an utterance is assumed to represent the speaker's emotion state at a particular point in time (analogous to sampling instances). The sequence of utterance emotion scores for a user forms their temporal emotion arc.[3]

Word-Count based Window: Here, the emotion score at a point in time is computed for a window of words (say, 100 words) that are uttered around

[2]Table 1 shows key details of the filtered TwitterSTMHD and Reddit eRisk datasets.

[3]The frequency and time of posting often differs between users, but we ignore that for now.

that point, and the window is moved forward by a fixed step size (say, 1 word at a time) to obtain the emotion score for the next time step. In prior work on constructing emotion arcs from temporally-ordered text, such sliding windows are usually employed to ensure smoother arcs that more accurately capture the flow of emotions over time.

Teodorescu and Mohammad (2023) conducted extensive quantitative evaluations of several hyperparameters involved in emotion arc construction, on datasets from diverse domains (including tweets) annotated with emotion scores. We follow many of their recommendations to construct emotion arcs for the utterances of each of our users. The texts are tokenized using the twokenizer[4] package to obtain a similarly-ordered sequence of words. Emotion scores are computed with window sizes of 100 words and 500 words each, and the window is moved forward by one word at each timepoint to obtain a series of overlapping emotion scores.

Emotion scoring method: Keeping in mind the necessity of an interpretable method of emotion scoring, we use word–emotion lexicons to compute the emotion scores of text spans. For each window, the emotion scores of its constituent words are averaged to obtain the window-level score for that emotion. Teodorescu and Mohammad (2023) showed that emotion arcs constructed with lexicon-based scoring methods, when used with sliding window sizes of 100 instances or more, can mimic the ground-truth emotion arcs with an accuracy of 0.9 or more.

Word–emotion scores are obtained from the NRC Emotion Intensity Lexicon, which associates close to 10,000 English words with a real-valued score between 0 and 1 for each dimension. A score of 0 indicates that the word has little to no association with that particular emotion, and a score of 1 indicates a high association.

Qualitative Checks on Emotion Lexicons: Lexicon-based methods for constructing emotion arcs are reliable and interpretable; however, it is good practice to modify the lexicon to the specific domain of use, in order to account for terms that are expected to be used in the target domain in a sense different from the predominant word sense (Mohammad, 2023).

We identify and remove words and bigrams whose usage on Twitter (and sometimes more colloquially) is markedly different from the predominant word sense annotated in the lexicons, such as *like* and *chaotic evil*. We also remove words and bigrams that are explicitly associated with mental health, such as *anxious*, *disorder* and *panic attack*. Though our EG metric does not explicitly rely on the presence of such terms to find associations with MHCs, we remove them in order to capture more fundamental differences in emotional expression between users in the MHC groups and the control group. The full list of stopwords is in Appendix B.

Hyperparameters: We additionally make the following choices of hyperparameters for constructing and comparing a pair of emotion arcs:

For a given pair of emotions, we drop all emotion terms that are common to the two lexicons before constructing their emotion arcs. This ensures that we are not using words associated with both emotions, giving us a clearer indication of co- endorsement. An utterance (or window) with no emotion terms from a particular emotion lexicon is assigned a score of 0. An alternative is to assign them a score of nan, in which case they are not considered a part of the emotion arc.[5]

4.2 Quantifying Emotion Granularity

We compute the emotion granularity metric as the negative of the Spearman correlation between each pair of emotions arcs, for each user.[6]

For each person, we average the correlation scores between emotion pairs in the different valence groups to obtain the following measures of emotion granularity (EG):

[4] https://github.com/myleott/ark-twokenize-py

[5] A visualization of the emotion arcs obtained using the utterance- level window for a sampled user from the Twitter-STMHD dataset is presented in Appendix E.

[6] A high correlation between two arcs indicates that the speaker is consistently and repeatedly expressing the two emotions concurrently; we hypothesize that this is an indicator of a lower ability to differentiate between the two emotions, and therefore a lower emotion granularity.

$EG(pos)$: The negative of (i.e., -1 times) the average of the correlation scores between each of the pairs of emotions in

$EG(neg)$: The negative of the average of the correlation scores between each of the pairs of emotions in the negative val

$EG(var)$: The negative of the average of the correlation scores between each of the pairs of emotions in the variable val

$EG(overall)$: Overall emotion granularity, measured as the negative of the average of the correlation scores between emotio

$EG(cross)$: Emotion granularity of crossgroup emotion pairs. That is, the negative of the average of the correlation score

We consider $EG(overall)$ to be the bottom line measure of emotion granularity for a user (analogous to that used in psychology studies). Note that cross-group pairs are not included in this measure.

5 Emotion Granularity and Mental Health

We now test if there are significant differences between the emotion granularities of each of the MHC groups and the control group, using t- tests. We first limit the users in each group by placing thresholds on (a) the number of user tweets with a valid emotion score (set to a minimum of 50), and (b) the number of unique lexicon terms used in their tweets (set to a minimum of 25). These thresholds ensure that we are drawing inferences based on users with valid emotion arcs, with sufficient lexicon coverage and temporal information.

We performed independent t- tests to compare emotion granularities between each of the MHCs and the control group, for each emotion group, using the SciPy library (Virtanen et al., 2020). To correct for multiple comparisons (eight tests performed for each MHC per emotion granularity group), we used the Benjamini- Hochberg procedure in the statsmodel's library (Seabold and Perk- told, 2010). Further details on the data assumptions for t- tests are in Appendix C.

5.1 Term Specificity as a Control

Lower emotion granularity occurs when, for a person, the concepts of the relevant emotions are so broad (and non- specific) that their meanings overlap substantially. This work is testing the hypothesis of whether people who have self- disclosed as having an MHC have lower emotion granularity than those that do not. However, another plausible hypothesis is that people in a particular group (e.g., MHC or the control) tend to use more specific words overall. Doing so would imply a higher specificity (i.e, a higher granularity) in their usage of all words, and that the high granularity of emotion words is simply a by-product of their general style of speaking (or posting online).

To ensure that the level of word specificity does not differ between MHCs and the control group and act as a confounder for our measure of emotion granularity, we performed a control experiment. We compute the average information content of the noun and verb terms in the posts of users in each group, and use this as a measure of the specificity of their language. We use the metric proposed in Resnik (1995), and implemented in the NLTK WordNet library, which combines information about the depth of the term in the WordNet tree hierarchy and its frequency of occurrence in a large corpus (here, the Brown corpus) to compute an information content score (Miller, 1995). We then compute the following measures of term specificity for each user:

$IC(n)$: The information content score for all nouns is averaged across

$IC(v)$: The information content score for all verbs is averaged across

Statistical tests for significant differences are similarly performed as described above (Section 5).

6 Results

In Table ?? we report the statistical results from the pairwise comparisons between each MHC and the control group, for the control experiment on general term specificity as well as emotion granularity, when scores are computed at the utterance- level.

All statistically significant differences between an MHC and the control group are described as either 'higher' or 'lower', and a dash (-) for no statistical difference. A 'lower' value in a cell indicates that the MHC (rows) has lower emotion granularity (or lower term specificity) than the control group, i.e.,

Table 2: The difference in emotion granularity between each MHC group and the control. A significant difference is indicated by the word 'lower' or 'higher', indicating the direction of the difference in granularity.

Dataset, MHC-Control	IC(n)	IC(v)	EG(pos)	EG(neg)	EG(var)	EG(cross)	EG(overall)
Twitter-STMHD							
ADHD-control	-	-	lower	lower	lower	lower	lower
Anxiety-control	-	-	lower	lower	lower	lower	lower
Bipolar-control	-	-	lower	lower	lower	lower	lower
MDD-control	-	-	lower	-	-	lower	lower
OCD-control	-	-	lower	lower	lower	lower	lower
PPD-control	-	-	-	lower	-	-	-
PTSD-control	-	-	lower	lower	lower	lower	lower
Depression-control	-	-	lower	lower	lower	lower	lower
Reddit eRisk							
Depression-control	-	-	lower	lower	-	lower	lower

higher correlation between emotion pairs in that group (columns); 'higher' indicates the MHC has higher emotion granularity (or higher term specificity) than the control group (i.e., lower correlation between emotion pairs in that group). In Table 12 in the Appendix, we also report the absolute Spearman correlation scores for each group. Below we summarize the results for each column.

6.1 Emotion Granularity as an Indicator of MHCs

IC(n) and IC(v): We do not see any significant differences in the information content of noun and verb terms ($IC(n)$ and $IC(v)$) between MHCs and the control group. This indicates that no group tends to use more specific or less specific language in general when posting on these platforms. Details on the statistical results are shown in Appendix H. **EG(pos):** All MHCs except for PPD had significantly lower positive emotion granularity than the control group (which had similar granularity compared to the control group). That is, tweeters in these MHC groups (ADHD, Anxiety, etc.) consistently expressed multiple positive emotions concurrently, more so than the control group.

EG(neg): All MHCs except MDD had significantly lower negative emotion granularity than the control group, in both datasets. Thus, tweeters in these MHCs were generally not differentiating between the negative emotions of anger, disgust, fear, and sadness, as well as the control group.

EG(var): Tweeters in the ADHD, Anxiety, Bipolar, OCD, PTSD, and Depression (Twitter- STMHD) had significantly lower variable emotion granularity than the control group (i.e., these groups generally differentiated between surprise and anticipation less than the control group).

EG(overall): All MHCs except PPD had significantly lower emotion granularity for emotion categories that express the same valence (the mixed valence emotions of surprise and anticipation are also included here). Tweeters in these groups are therefore expressing multiple close emotions frequently with one another - more so than the control group.

EG(cross): All MHCs except PPD had significantly lower granularity between emotion pairs that come from different valence groups. This indicates that positive and negative emotions are expressed together more frequently by tweeters in these groups compared to the control, as well as emotions like joy (positive valence) and surprise (variable valence).

Discussion: These results demonstrate that our measures of emotion granularity from text are consistently lower for users in the MHC groups compared to the control. The term specificity results also tell us that it is the specificity of emotion word usage in particular that is differentiating MHCs from the control group.

Aligning with self- report studies in psychology, the emotion granularity between negative- valence emotions is lower for most (7 out of 8) MHCs in our datasets with utterance- level operationalizations. Positive emotion granularity is also correlated with many of the MHCs (7 out of 8 disorders). In general, the granularity of emotional expression between within- group emotion pairs is lower for all MHC groups compared to the control in both datasets, except PPD. This is in line with both the theoretical and conceptual links established in the psychology literature on emotion granularity and mental health: the ability to better differentiate between emotion concepts that are close to one another leads to more adaptive health behaviour.

While emotion pairs from differently- valenced emotion groups are not usually operationalized in affective science experiments, we find that this measure is also significantly lower in many MHCs. Further investiga-

tions into what the concurrent expression of positive and negative emotions means, for emotion granularity and emotion dynamics in general, are interesting research directions.

Variation with hyperparameter choices: We observe only minor variations from the results reported in Table ??, when the hyperparameters described in Section 4.1 for emotion arc construction were changed - less than 10% of the cells differed in their values across all variations. We provide a more detailed report in Appendix F.

6.2 Additional Window Sizes

We also examined how the measures of differences in emotion granularity between MHCs and the control change when we compute emotion scores with larger window sizes.

Many of the utterance- level outcomes are replicated for negative, positive, and overall emotion granularity with window sizes 100 and 500. Some measures are no longer significantly different between certain MHCs and the control. We also find that $EG(cross)$ is higher for certain MHCs (Anxiety, PPD, PTSD, Depression in Twitter- STMHD) when compared to the control, i.e., users in the control group are expressing negative and positive emotions together more frequently than those in the MHCs.

With larger window sizes, we end up capturing emotions expressed by the individual over longer time spans (tweets posted over the span of several hours or days), rather than co- endorsement at the same time. We hypothesize that these effects of dialecticism, where the control group has a higher co- occurrence of cross- valence group emotions, are capturing the extent to which users balance negative emotions with positive emotions (and vice versa). The consistent effects with 100 and 500- word windows, and for several MHCs, makes this a promising area for future work. All emotion granularity measures with window sized 100 and 500 are reported in Appendix F.1.

6.3 Individual Emotion Pairs

In order to understand which emotion pairs are expressed together more frequently (resulting in lower emotion granularity), we performed the same significance tests as before between MHCs and the control for correlation scores between all individual emotion pairs. We found that:

Seven out of the eight MHCs in the TwitterSTMHD dataset had a lower granularity (a higher correlation) for anger- disgust (except PPD) and anger- sadness (except MDD) in the negative valence group. All eight MHCs had a lower emotion granularity (higher correlation) between multiple cross- group emotion pairs, notably those involving the mixed- valence emotions of anticipation and trust. Contrary to trends, the Bipolar group had a higher emotion granularity (i.e., a lower correlation of emotion arcs) for the cross-group emotion pairs of anger- joy and fear- joy.

Detailed results for each of the emotion pairs and all MHCs are in Appendix G, Table 10.

Discussion: While lower granularity among certain emotion pairs consistently function as indicators of all MHCs, we also see a few instances where MHCs (specifically Bipolar disorder) have a higher granularity between the emotions when compared to the control. These findings are of interest to researchers studying the links between how emotions are expressed in text, and how they vary with different MHCs.

7 Conclusion

In this work, we operationalized for the first time a computational measure of emotion granularity that can be derived from the textual utterances of individuals. We applied this measure to two social media datasets of posts from individuals who have self- disclosed as having an MHC. Our findings showed that our measure of negative emotion granularity is significantly lower for 7 out of the 8 MHC groups under consideration when compared to a control group, at an aggregate- level. Also, all MHCs except for PPD had lower overall emotion granularity (and lower positive emotion granularity) compared to the control group. Our work makes an important contribution towards deriving aggregate- level indicators of emotional health from the large amounts of utterance data available on social media platforms. We hope

this opens up an avenue of future work to explore emotional expression in text and mental health.

8 Limitations

Our work uses the social media utterances of individuals to derive measures of emotional expression that, at an aggregate level, are found to correlate with multiple mental health conditions. While we use datasets that were compiled by other researchers in the field, we stress that they may not be representative of the general population. Our methods therefore cannot be directly applied to make inferences on other datasets without a careful experimental validation first. The datasets we study rely on self- disclosures made on social media platforms; it is possible that users report only one such MHC but are diagnosed with others, or that they misrepresent their diagnoses. Further, the users in the control groups may include those who have chosen to simply not self- disclose on these platforms. This can occur due to many reasons, like social desirability (Latkin et al., 2017) or impression management (Tedeschi, 2013). Nonetheless, since we draw inferences at an aggregate level, the methods used can overcome some amount of noisy data.

The set of emotions that we have considered in our measurement of emotion granularity are also limited to those for which we can computationally obtain text-derived emotion scores. These eight emotions do not represent the wide range of emotion concepts that exist and are experienced and expressed by us with language, and future research can attempt to expand our operationalization to more emotion concepts. It should be noted though, that past psychology studies on emotion granularity have also tended to explore small sets of emotions, largely because it is cumbersome to ask users about how they feel for a large set of emotions.

The emotion lexicons that we use are some of the largest that exist with wide coverage and large number of annotators (thousands of people as opposed to just a handful). However, no lexicon can cover the full range of linguistic and cultural diversity in emotion expression. The lexicons are largely restricted to words that are most commonly used in Standard American English and they capture emotion associations as judged by American native speakers of English. See Mohammad (2023) for a discussion of the limitations and best- practises in the use of emotion lexicons.

Lastly, further work should explore if the relationships we found hold around various social factors such as age, region, language, etc. As we focus on English text, and the region of users is not known (some information could be extracted from user profiles in the Twitter- STMHD dataset however it is fairly noisy), conclusions should be drawn cautiously across various sociolinguistic factors.

9 Ethics Statement

Our approach, as with all data- driven models of determining indicators of mental health, should be considered as aggregate- level indicators, rather than biomarkers for individuals (Guntuku et al., 2017). We do not attempt to predict the presence of MHCs for individual users at any stage of the process. These measures should also not be taken as standalone indicators of mental health or mental wellness, even at the population level, but rather as an additional metric that can be used in conjunction with other population- level markers, and with the expertise of clinicians, psychologists, and public health experts.

Individuals vary considerably in how, and how well, they express their internal emotional states using language. Our method of assessing the emotional states of users based on their utterances may miss several linguistic cues of emotion expression, and may not account for individual variation or the extent to which these emotions are expressed on social media. The emotionality of one's language may also be conveying information about the emotions of the speaker, the listener, or something or someone else mentioned in the utterances. See further discussions of ethical considerations when using computational methods for affective science in Mohammad (2023, 2022).

A Twitter-STMHD Dataset

Suhavi et al. (2022) created a regular expression pattern to identify posts which contained a self- disclosure of a diagnosis and the diagnosis name (using a lexicon of common synonyms, abbreviations, etc.) such as 'diagnosed with X'. They collected a large set of tweets using the regex. This resulted in a preliminary dataset of users with potential MHC diagnoses. To handle false positives (e.g., 'my family member has been diagnosed with X', or 'I was not diagnosed with X'), the dataset was split into two non- overlapping parts, one of which was manually annotated, and the other using an updated and high- precision regex. In the part that was annotated by hand, each tweet was annotated by two members of the team. A user was only included in the dataset if both annotations were positive as self- disclosing for a particular class. A licensed clinical psychologist found the 500- tweet sample to be 99.2% accurate. The manual annotations were used to refine the regular expressions and diagnosis name lexicon. This updated search pattern was applied to the other dataset split. To verify the quality of the updated regex, the authors applied it to the manually annotated dataset split. When considering the manual annotations as correct, the regex was found to be 94% accurate.

B Lexicon Words Removed

We considered the following sets of terms to be stopwords, which do not contribute to the emotion score of an utterance, for our analysis:

Common stopwords: We remove common English stopwords, such as the, of, for, etc. We use the list of English stopwords from the Python NLTK library. The full list can be found at https://gist.github.com/sebleier/554280. **Domain- specific stopwords:** We remove terms (words and word pairs) whose dominant usage on social media platforms differs from their annotated sense (e.g, like, chaotic evil, good morning). The full list of these terms is in Table **??**. **MHC- associated terms:** Finally, we filter out terms that are explicitly associated with the MHCs that we consider, such as anxiety, mental health, and panic attack. The full list of terms is in Table **??**.

Table 3: Twitter-specific words and bigrams removed from the emotion lexicons.

love, flu shot, raptor, discord, christmas, good day, good morning, good evening, birthday, good night, good afternoon, bloody murder, pretty, true crime, full time, gut punch, vibe, wholesome content, slur word, life time, vote, jump scare, shot chocolate, chaotic evil, trump, fever dream, chaotic energy, chaotic good, like, guilty pleasure, chaotic neutral, hot mess

Table 4: Mental health specific terms removed from the emotion lexicons.

disability, ptsd, psychosis, adhd, suicide, depressive, depressed, disorder, anxiety, mental health, anxious, mental illness, disabled, panic attack

C Statistical Assumptions

Below we describe in more depth the requirements for performing an independent t- test, which was done in our analyses.

The dependent variable must be measured using a continuous scale: emotion granularity is measured as the average of Spearman correlation between emotion arcs in the group, resulting in continuous values. The independent variable must have two categorical and independent groups: Our independent variable is diagnosis, which is either an MHC or the control group. Independence of observations: Since the text stream utterances come from different people, we can assume these are independent observations. Approximately normally distributed dependent variable for each group of independent variable: Given the large number of people and number of utterances per person in our dataset, we can assume that the means of the data for each group is approximately normally distributed according to the law of large numbers. Further, the t- test is robust to violations of normality. Homogeneity of variance: We performed Levene's

test for homogeneity of variance to verify whether this assumption is met. Our data did not meet this assumption, therefore we performed t- tests with the unequal variance setting as True in SciPy.

D Emotion Lexicons

In Table 5, we report statistics on the number of emotion terms in each lexicon for the eight emotions we consider in this work, and the number of terms common to and mutually- exclusive between each emotion pair.

E Visualization of Emotion Arcs

In Figure 1, we plot the emotion arcs for the fear-sadness emotion pair, from the negative valence group, for a tweeter from an MHC group of the Twitter-STMHD dataset. Emotion scores are computed and plotted at the utterance- level, i.e, independently for each tweet by the user. Note that larger window sizes and overlapping windows will lead to smoother arcs.

IMAGE NOT PROVIDED
Visualization of emotion arcs for fear-sadness pair (utterance-level).

Figure 1: Emotion arcs for fear and sadness for a sample user.

F Emotion Granularity: Hyper-parameters

We report the results of the statistical analyses of emotion granularity when emotion arcs were generated using different choices of the hyperparameters described in Section 4.1. Table **??** reports the results when non- lexicon terms (and tweets) are assigned a score of 0, and only mutually- exclusive emotion terms are considered, similar to Table 2, but no user thresholds on number of tweets and unique emotion terms are applied. Table **??** reports the results when non- lexicon terms (and tweets) are not considered, and user thresholds are set to 50 and 25, (similar to Table 2), and only mutually- exclusive emotion terms are considered (similar to Table 2).

We find that largely the results do not change. However, when non- lexicon terms and tweets are ignored, this results in a smaller set of tweets to compute the emotion arc over, and fewer tweeters who meet the user thresholds for each group. This results in signals turning off for certain MHCs.

Table 5: Emotion Granularity - hyperparameter variations: The difference in emotion granularity between each MHC group and the control. A significant difference is indicated by the word 'lower' or 'higher', indicating the direction of the difference in granularity. Non-lexicon terms and tweets are assigned a score of zero; user tweet and unique term thresholds are both set to 0, and only mutually-exclusive emotion terms are considered.

Dataset, MHC-Control	IC(n)	IC(v)	EG(pos)	EG(neg)	EG(var)	EG(cross)	EG(overall)
Twitter-STMHD							
ADHD-control	-	-	lower	lower	lower	lower	lower
Anxiety-control	-	-	lower	lower	lower	lower	lower
Bipolar-control	-	-	-	lower	lower	lower	lower
MDD-control	-	-	lower	-	-	lower	lower
OCD-control	-	-	lower	lower	lower	lower	lower
PPD-control	-	-	-	lower	lower	-	lower
PTSD-control	-	-	lower	lower	lower	lower	lower
Depression-control	-	-	lower	lower	lower	lower	lower
Reddit eRisk							
Depression-control	-	-	lower	lower	lower	lower	lower

Table 6: Emotion Granularity - hyperparameter variations: The difference in emotion granularity between each MHC group and the control. A significant difference is indicated by the word 'lower' or 'higher', indicating the direction of the difference in granularity. Non-lexicon terms and tweets are discarded; user tweet and unique term thresholds are set to 50 and 25, and only mutually-exclusive emotion terms are considered.

Dataset, MHC-Control	IC(n)	IC(v)	EG(pos)	EG(neg)	EG(var)	EG(cross)	EG(overall)
Twitter-STMHD							
ADHD-control	-	-	lower	lower	lower	lower	lower
Anxiety-control	-	-	-	lower	lower	lower	lower
Bipolar-control	-	-	lower	lower	lower	lower	lower
MDD-control	-	-	-	-	-	-	-
OCD-control	-	-	lower	lower	lower	lower	lower
PPD-control	-	-	lower	lower	lower	upper	lower
PTSD-control	-	-	lower	lower	lower	lower	lower
Depression-control	-	-	higher	lower	lower	lower	-
Reddit eRisk							
Depression-control	-	-	-	lower	lower	lower	lower

F.1 Various Window Sizes

We report the results of the statistical analyses of emotion granularity when emotion arcs were generated using two other window sizes: 100 (Table **??**) and 500 (Table **??**). All other hyperparameters are the same as for Table 2.

We find that largely the results do not change, however there are some differences in the scenario when the dataset was smaller (e.g., eRisk dataset or MHC such as MDD). In such cases, when the window size is increased, it is possible that several emotional experiences occurred, resulting in a weaker signal of emotion granularity.

Table 7: Emotion Granularity - using window 100: The difference in emotion granularity between each MHC group and the control. A significant difference is indicated by the word 'lower' or 'higher', indicating the direction of the difference in granularity.

Dataset, MHC-Control	IC(n)	IC(v)	EG(pos)	EG(neg)	EG(var)	EG(cross)	EG(overall)
Twitter-STMHD							
ADHD-control	-	-	lower	lower	higher	lower	lower
Anxiety-control	-	-	lower	lower	lower	higher	lower
Bipolar-control	-	-	lower	lower	-	-	lower
MDD-control	-	-	lower	lower	-	-	-
OCD-control	-	-	lower	lower	-	lower	-
PPD-control	-	-	-	-	-	lower	higher
PTSD-control	-	-	-	lower	-	higher	lower
Depression-control	-	-	lower	lower	higher	higher	lower
Reddit eRisk							
Depression-control	-	-	lower	-	lower	-	lower

Table 8: Emotion Granularity - using window 500: The difference in emotion granularity between each MHC group and the control. A significant difference is indicated by the word 'lower' or 'higher', indicating the direction of the difference in granularity.

Dataset, MHC-Control	IC(n)	IC(v)	EG(pos)	EG(neg)	EG(var)	EG(cross)	EG(overall)
Twitter-STMHD							
ADHD-control	-	-	lower	lower	-	lower	lower
Anxiety-control	-	-	lower	lower	higher	higher	lower
Bipolar-control	-	-	lower	lower	-	-	lower
MDD-control	-	-	-	-	-	-	lower
OCD-control	-	-	lower	-	-	-	-
PPD-control	-	-	-	-	-	-	higher
PTSD-control	-	-	-	lower	-	higher	lower
Depression-control	-	-	lower	lower	higher	higher	lower
Reddit eRisk							
Depression-control	-	-	lower	-	-	-	-

G Emotion Granularity: Emotion Pairs

In Table **??** we report the pairwise emotion granularity results when testing for significant differences between MHCs and the control group.

Table 9: Emotion Granularity - emotion Pairs: The difference in emotion granularity between each emotion pair, for each MHC group and the control in the Twitter-STMHD dataset. A significant difference is indicated by the word 'lower' or 'higher', indicating the direction of the difference.

Emotion Pair	ADHD	Anxiety	Bipolar	Depression	MDD	OCD	PPD	PTSD
anger-anticipation	lower	lower	-	lower	-	lower	-	lower
anger-disgust	lower	lower	lower	lower	lower	lower	-	lower
anger-fear	lower	lower	lower	lower	-	lower	lower	lower
anger-joy	lower	lower	higher	lower	-	lower	-	-
anger-sadness	lower	lower	lower	lower	-	lower	lower	lower
anger-surprise	lower	lower	-	lower	-	lower	-	lower
anger-trust	lower	lower	lower	lower	lower	lower	-	lower
anticipation-disgust	lower	lower	lower	lower	-	lower	-	lower
anticipation-fear	lower	lower	-	lower	lower	lower	lower	lower
anticipation-joy	lower	lower	lower	lower	lower	lower	lower	-
anticipation-sadness	lower	lower	lower	lower	-	lower	-	lower
anticipation-surprise	lower	lower	lower	lower	-	lower	-	lower
anticipation-trust	lower	lower	lower	lower	lower	lower	lower	-
disgust-fear	lower	lower	lower	lower	-	lower	lower	lower
disgust-joy	lower	lower	-	lower	-	lower	-	lower
disgust-sadness	lower	lower	lower	lower	-	lower	lower	lower
disgust-surprise	lower	lower	lower	lower	-	lower	-	lower
disgust-trust	lower	lower	lower	lower	lower	lower	lower	-
fear-joy	lower	lower	higher	lower	-	lower	-	lower
fear-sadness	lower	lower	lower	lower	-	lower	lower	lower
fear-surprise	lower	lower	lower	lower	-	lower	-	lower
fear-trust	lower	lower	lower	lower	lower	lower	lower	-
joy-sadness	lower	lower	-	lower	-	lower	-	lower
joy-surprise	lower	lower	-	lower	-	lower	-	lower
joy-trust	lower	lower	lower	lower	lower	lower	-	lower
sadness-surprise	lower	lower	lower	lower	-	lower	lower	lower
sadness-trust	lower	lower	lower	lower	lower	lower	lower	-
surprise-trust	lower	lower	-	lower	-	lower	lower	lower

H Term Specificity Results

Table **??** shows the results of the term specificity experiments described in Section 5.1 measuring information content. For both nouns and verbs, none of the diagnoses had significantly different term specificity levels compared to the control group in both the Twitter- STMHD and eRisk datasets. This verifies that the significant differences between the MHCs and the control group for emotion granularity is not due to varying word specificity levels in these groups.

Table 10: Information Content: The degrees of freedom, t-statistic and p- value for the word specificity experiments described in Section 5.1.

Dataset	MHC-Control	POS	df	T-Statistic	P-value
Twitter-STMHD	ADHD-control	Noun	2368.64	-1.94	0.144
	Anxiety-control		2233.36	-0.58	0.718
	Bipolar-control		1726.26	-2.40	0.131
	MDD-control		226.83	-0.52	0.718
	OCD-control		1817.93	1.93	0.144
	PPD-control		178.33	-0.49	0.718
	PTSD-control		2237.82	-0.54	0.718
	Depression-control		2245.64	-0.27	0.787
	Depression-control		128.95	0.98	0.330
Twitter-STMHD	ADHD-control	Verb	2248.45	-0.73	0.530
	Anxiety-control		2235.85	1.12	0.420
	Bipolar-control		1852.44	-2.10	0.096
	MDD-control		213.0	1.36	0.354
	OCD-control		1645.17	2.28	0.091
	PPD-control		169.49	0.98	0.438
	PTSD-control		2351.12	2.53	0.091
	Depression-control		2274.59	0.54	0.589
	Depression-control		110.40	1.59	0.116

I Emotion Correlations

Table **??** shows the group- averaged Spearman correlations for emotion pairs in the positive, negative, mixed valence groups, and the within- group and cross-group averages, for the Control groups, and the delta from these values for each MHC in both datasets.

Table **??** shows the Spearman correlation between emotion arcs for all pairs of emotions for the control group. These results indicate that as baselines largely emotions in the same group (e.g., positive, negative, mixed, overall) co- occur more often than emotions across groups.

Table 11: Emotion Granularity - Spearman correlations: Spearman correlation values between utterance-level emotion arcs for the Control group, and the delta for each MHC when compared to the Control group. Emotion granularity is defined as the negative of these correlations (i.e, higher correlations imply a lower granularity). Hyperparameters are the same as in Table 2.

Dataset, MHC	EG(pos)	EG(neg)	EG(var)	EG(Cross)	EG(overall)
Twitter-STMHD					
Control	0.027	0.023	0.012	0.006	0.022
ADHD	-0.012*	-0.008*	-0.005*	-0.010*	-0.010*
Anxiety	-0.012*	-0.008*	-0.003*	-0.008*	-0.010*
Bipolar	-0.004*	-0.008*	-0.004*	-0.002*	-0.006*
MDD	-0.011*	-0.002	-0.001	-0.005*	-0.005*
OCD	-0.013*	-0.009*	-0.006*	-0.009*	-0.009*
PPD	-0.005	-0.009*	-0.005	-0.004	-0.003
PTSD	-0.013*	-0.014*	-0.006*	-0.008*	-0.013*
Depression	-0.011*	-0.005*	-0.003*	-0.005*	-0.008*
Reddit eRisk					
Control	0.114	0.117	0.090	0.094	0.112
Depression	-0.016*	-0.021*	-0.012	-0.022*	-0.017*

Table 12: Emotion-Emotion Spearman correlations: Spearman correlation values between pairs of utterance-level emotion arcs for the all users in the control group of the Twitter-STMHD dataset. Hyperparameters are the same as in Table 2.

	anger	anticipation	disgust	fear	joy
anger	–	0.003	0.020	0.027	-0.010
anticipation	–	–	0.003	0.007	0.02
disgust	–	–	–	0.023	-0.01
fear	–	–	–	–	-0.00
joy	–	–	–	–	–
sadness	–	–	–	–	
surprise	–	–	–	–	