What type of happiness are you looking for? - A closer look at detecting mental health from language

Alina Arseniev-Koehler[1]✉, Sharon Mozgai[2] and Stefan Scherer[2]

[1]University of California, Los Angeles, CA; arsena@g.ucla.edu
[2]USC Institute for Creative Technologies, Playa Vista, CA

Abstract

Computational models to detect mental illnesses from text and speech could enhance our understanding of mental health while offering opportunities for early detection and intervention. However, these models are often disconnected from the lived experience of depression and the larger diagnostic debates in mental health. This article investigates these disconnects, primarily focusing on the labels used to diagnose depression, how these labels are computationally represented, and the performance metrics used to evaluate computational models. We also consider how medical instruments used to measure depression, such as the Patient Health Questionnaire (PHQ), contribute to these disconnects. To illustrate our points, we incorporate mixed-methods analyses of 698 interviews on emotional health, which are coupled with self-report PHQ screens for depression. We propose possible strategies to bridge these gaps between modern psychiatric understandings of depression, lay experience of depression, and computational representation.

1 Introduction

Valid, reliable tools to automatically detect mental illness from text and speech would be groundbreaking. Such tools could provide new opportunities for early detection and intervention in combination with clinician opinions. They would also open new doors for research to expand our still nascent understanding of the causes and mechanisms of mental health. The prospect of such tools have inspired a burgeoning area of research on detecting mental health.

Given prevalence and heavy toll of depression, it may not be surprising that this mental illness the focus of modeling efforts (De Choudhury et al., 2013; Fraser et al., 2016; Gupta et al., 2014; Resnik et al., 2013; Williamson et al., 2016; Schwartz et al., 2014; Howes et al., 2014; Fraser et al., 2016; Tsugawa et al., 2015; Nguyen et al., 2014; De Choudhury et al., 2014; Tsugawa et al., 2015; Nadeem, 2016; Reece et al., 2017; Guntuku et al., 2017; De Choudhury et al., 2016). Depression is characterized by low mood, a lack of interest, cognitive and psychomotor impairment, and suicidal ideation. And, nearly one in five Americans will experience depression at some point in their lifetimes (Kessler and Bromet, 2013).

Such models report compelling accuracy rates at detecting depression from written and transcribed verbal data. Many of these modeling efforts cite a long-term common vision of an end-to-end, automated system which may even be deployable in clinical settings. However, computational models of depression are often disconnected from the lived depression experience and siloed from larger debates on how to characterize and classify mental health. Indeed, characterizing and diagnosing depression is an ongoing, active area of debate fueled by nearly a century of clinical research (Bowins, 2015; Insel et al., 2010; Insel, 2013). Meanwhile, laypeople and those actually experiencing depression construct their own meanings of this mental illness (Karp, 2016).

This paper re-examines the detection of depression from language, and revisits old and current debates in mental health classification. Along the way, we highlight strengths and weaknesses of modeling approaches and propose several strategies for more reflexive modeling.

2 Methods and Data

Primarily, we review peer-reviewed research detecting and predicting depression from text data. Importantly, we are specifically interested in efforts to detect depression from written text data or transcribed verbal data, rather than vocal fea-

Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, pages 1–12
New Orleans, Louisiana, June 5, 2018.

tures. Patterns of vocal features are better understood (Cummins et al., 2015), and text evidence is a promising modality for depression detection (Calvo et al., 2017). Further, these two modalities are different in that language is a primary medium by which we create and communicate meaning, and this is often done very consciously (Blumer, 1986). We focus on models detecting and predicting depression, but incorporate ideas from modeling other mental illnesses and emotions.

We provide additional quantitative and qualitative evidence from ongoing analyses of interviews with 698 participants from the Distress Analysis Corpus (DAIC) (Gratch et al., 2014). Participants are drawn from two populations living in the greater Los Angeles. First, the general public and second, veterans of the U.S. armed forces. These interviews are conducted with an avatar, Ellie, and are intended to simulate clinical interviews screening for mental health symptoms (DeVault et al., 2014). Interviews were automatically transcribed with IBM Watson; thus simulating how an end-to-end system for mental health screening from verbal data might work. Interviews are coupled with self-report measures on psychological health, such as an 8-item version of the Patient Health Questionnaire (PHQ-8) (Kroenke and Spitzer, 2002).

The PHQ is a clinically validated, self-administered screen for depression to capture symptoms of depression according to the Diagnostic and Statistical Manual of Mental Disorders (DSM). Abbreviated, validated versions of the PHQ are commonly used, particularly an 8 item version (PHQ-8). Briefly, the eight items in the PHQ-8 include the symptoms: 1) changes in appetite 2) feelings of failure or worthlessness 3) tiredness or lethargy 4) trouble sleeping 5) trouble concentrating 6) lack of interest or ability to take pleasure 7) depressed mood, and 8) psychomotor impairment, such as fidgeting or moving slowly. A nine item version (PHQ-9) is also commonly used, which includes a ninth item regarding suicidal ideation. Possible scores on the PHQ-8 ranges from 0-24, and individuals with scores of 10 or greater are considered as currently having depression. The mean PHQ-8 score among our participants was six, and 175 (25%) of our participants scored as currently having depression according to this scale.

For a pilot set of 140 of these participants, we obtained layperson annotations of participants' mental health. Specifically, we asked crowd workers to read excerpts of de-identified, transcribed interview data, and then rate how likely they thought a speaker had depression based on the transcribed utterances. Response options were "very unlikely," "unlikely," "likely," and "very likely," or that there was "no evidence" either way for depression. Crowd workers were asked to repeat this task for eight symptoms according to the PHQ-8 list of symptoms. We use 100-word excerpts to balance having enough content with having granular labels. The 140 participants' transcripts yielded 1523 unique utterances, each of which were rated by three different crowd workers for a total of 4569 rated utterances.

Qualitative analyses included: 1) for a subset of interviews, open-coding entire interviews for how participants talk about mental health and emotions and 2) searching all interviews for lexicon relevant to depression (e.g., depressed, depression, depressing, sad, sadness, blue, happy, happiness, content) and then open-coding interview sections with this lexicon and comparing this data to interviewees PHQ-8 scores (Burnard, 1991). Incorporating qualitative data gives a voice to participants who have actually experienced emotional distress and reminds us of the human element behind quantitative representations. Qualitative analyses were performed by the first author. Quantitative analyses include data summaries, basic statistics, and inter-rater agreements for crowd workers' ratings. We use non-parametric statistics as needed, depending on data distributions.

3 Describing, Detecting, and Explaining

To detect mental health from text data, a set of handcrafted features is usually extracted and then fed into a supervised machine-learning classifier, such as a support vector machine (e.g. De Choudhury et al. 2013). Hand-crafted features commonly used include markers of linguistic style based on published dictionaries and depression lexicon (e.g., the use of "depressed," and "sad"). Topics derived from Latent Dirichlet allocation (LDA) topic models are also frequently used features (Blei et al., 2003).

A common dictionary for linguistic style and content is the Linguistic Inquiry and Word Count, or LIWC (Pennebaker et al., 2015). LIWC includes pyschometrically validated bag-of-words categories such as pronouns, tense, and lexicon

about emotions. Like many other hand-crafted features, LIWC offers explanatory power and transparency, and lends itself to hypothesis driven models for detecting depression. For example, individuals who are considered depressed tend to use words about negative emotions and first-person singular (e.g., "I") more often than those who are not considered depressed (Rude et al., 2004). We replicate these patterns in our data as well. Specifically, we find that those with higher PHQ-8 scores tend to use more words about negative emotions (Spearman ρ = .09, $p < .05$) and particularly sad emotions (Spearman ρ= .25, $p < .001$). Further, those with higher PHQ-8 scores tend to use more first-person singular pronouns (Spearman ρ= .13, $p < .001$) and fewer third-person singular pronouns (e.g.,"we") (Spearman ρ= -.11, $p < .001$). These patterns are thought to reflect that depression corresponds to negative thinking and to turning inward (Rude et al., 2004).

Of course, a model for detecting depression need not have features that are so carefully crafted or transparent. Indeed, modeling already often includes some dimensionality reduction step, such as Principal Component Analysis, on an abundance of features. Other more automated feature extraction from text data is less common in this realm but may be useful to find new features with strong predictive power, even if they do not have strong explanatory power (Shmueli, 2010). In most situations where a model to detect depression from language would be used, predictive power is more useful than explanatory insight. With enough data, methods such as long short-term memory (LSTM) neural networks may be promising ways to extract new and perhaps less explicit features, and account for higher level patterns in language, such as the order of words (Hochreiter and Schmidhuber, 1997).

4 What's in a label? Revisiting mental health labels in natural language processing

In predictive modeling and detection, labels are often treated as the objective truth. They are the gold-standard a model seeks to match, and against which errors are compared. This places tremendous confidence in these labels, particularly when labels are binary measures of mental wellness or illness. However, these labels have their own back-story in which they are created and re-created by clinicians, medical institutions, and researchers. Indeed, nearly a hundred years of research has produced modern screening for depression (Davison, 2006). Particularly in the realm of mental health, we can't take labels at face-value.

Most studies detecting depression use labels from self-report diagnostic scales, such as the PHQ. Implicitly, these scales are proxies for psychiatric ratings from structured interviews. Of course, self-report diagnostic scales are an imperfect proxy (Thombs et al., 2014). For example, in an original validation study for the PHQ-9, the PHQ-9 reaches 88% sensitivity compared to mental health professionals' ratings (Kroenke and Spitzer, 2002; Kroenke et al., 2001). Rates for the sensitivity of the PHQ-8 are more like 77% in subsequent validation studies (Arroll et al., 2010; Gilbody et al., 2007). Thus, even an algorithm which perfectly predicts PHQ scores from language, with tight confidence intervals on performance metrics, likely has a wide margin for errors for detecting depression when compared to a mental health professional rather than the proxy measure on which it is trained. The limitations of these diagnostic scales, and debates underlying them, are too often swept aside as we feed labels into algorithms.

A few studies detecting mental health from language use claims of diagnosis as a label for depression, such as *I was diagnosed with having P.T.S.D . . . So today I started therapy, she diagnosed me with anorexia, depression, anxiety disorder, post traumatic stress disorder and . . . .* (Coppersmith et al., 2014). On the one hand, given the stigma around mental illness and negative emotions, this approach risks missing those who do not to "come-out" as depressed. It also risks missing those who may do not share clinical meanings of depression, or are unaware they might have depression symptoms. On the other hand, this approach esteems an individuals' self-awareness and own experience of their mental health as the gold standard, reminiscent of a phenomenological approach to the depression experience. This label does not assume a form and structure for depression, unlike diagnostic scales. This lack of standardized definitions for depression makes comparison across settings and studies challenging. But, it also enables depression to be defined by the individuals own experience rather than an external scale or criteria.

Our data, too, shows that selecting a label is no easy task. For example, we find that some participants are categorized as low risk of depression according to the PHQ-8, when they openly talk in interviews about symptoms or about struggling with depression. One participant believes their best friend would describe them as happy, but scores nearly at the maximum value for depression on the PHQ-8. Another participant mentions *I can't even fathom happiness,* while reporting a PHQ-8 score just above the cutoff for mild depression: qualitatively and quantitatively these two reports tell different stories. Similar mismatches between patient stories and diagnostic scores have been noted by general practitioners (Davidsen and Fosgerau, 2014).

We see other types of mismatches between lived experiences of depression and quantitative representations of depression as well. For example, another participant in our study - who is not categorized as currently depressed based on the PHQ-8 - says, *yeah i've been diagnosed with depression once so i feel like it's one of those things that uh is something i have to keep in check throughout my entire life.* It is possible that this participant is not categorized as depressed precisely because they are successfully *managing* depression. Depression commonly recurs, and linguistic patterns of depression may vary across the trajectory of a depression experience (Capecelatro et al., 2013). Indeed, many labels, such as the PHQ, were originally intended to capture *current* depression episodes.

Some of the inconsistency between scores, feelings, and verbal expressions may also be due to the effects of social desirability and stigma in reporting mental health. In fact, Resnik et al. suggest "throwing out" participants who score 0 or 1 on such scales, as these individuals tend to report based on social desirability rather than a clear picture of their emotional health. In our data, however, this would constitute throwing out around a quarter of participants; 126 (19%) participants reported a 0, and 181 (27%) report a 0 or 1 on the PHQ-8.

So far in our discussion of labels for depression detection, we have presented psychiatric ratings as the comparison points for self-report measures on mental health. However, unlike a "broken bone," or a "sprained wrist," mental health is a gray area. Mental health is largely defined by our conceptions of what is "normal" and what is "disordered" — conceptions which change across culture and time (Karp, 2016).

4.1 What is mental illness, anyway? The myth of the gold standard

Mental illness manifests diversely across people, contexts, and cultures (Karp, 2016; Halbreich and Karkun, 2006; Canino and Alegría, 2008). Illnesses and symptoms are differently defined, but also differently expressed. For example, Chinese and Chinese-Americans tend to express Western definitions of depression more as somatic symptoms, rather than affective symptoms, while this pattern is reversed for Caucasians (Parker et al., 2001; Huang et al., 2006). Further, other cultures have categories for mental illness that we do not have in western thinking, such as the Japanese syndrome *taijin-kyofusho*, roughly translated as a "fear of interpersonal relations" (Tarumi et al., 2004). Whereas we carve out definitions like "depression," others may carve out different "idioms of distress" (Radden, 2003). Delineating labels for depression may be as much a cultural, as it is a medical, endeavor.

After a century of Western medical research, depression remains enigmatic in medicine and psychiatry (Davison, 2006). Diagnostic manuals, such as the DSM-V, were developed to enable reliable diagnosis by using precise definitions, criterion and nomenclature. They replaced phenomenological approaches to psychiatry, which focused on subjective experiences rather than than aiming to understand behavior by fitting it into preexisting definitions (Andreasen, 2006; Jacob, 2012; Mullen, 2006). In modern psychiatry, diagnoses are descriptive, co-occurring clusters of symptoms. They do not reference to underlying mechanisms or causes, and categories provide little information on treatment responses (Radden, 2003; Insel et al., 2010; Insel, 2013; Paykel, 2008). In the words of former director of the National Institute for Mental Heath (NIMH), Thomas Insel, *in the rest of medicine, this would be equivalent to creating diagnostic systems based on the nature of chest pain or the quality of fever* (2013). While psychiatric diagnostic manuals are intended for reliability — validity is their *weakness* (Insel, 2013).

Despite efforts to standardize diagnostic procedures, understandings of depression vary even among practicing medical professionals. For ex-

ample, unlike psychiatrists, general practitioners consider depression a gray area and doubt the utility of diagnostic tools (Davidsen and Fosgerau, 2014). And, even among psychiatrists, unreliability of depression diagnoses remains an well-documented issue (Aboraya et al., 2006). This issue may be even more pronounced when text is the only modality available for diagnostic clues. Resnik et al. examined interrater reliability of three psychologists who were asked to make depression diagnoses based on subjects' written text. These practicing psychologists were licensed and spend significant time in assessment and diagnosis of psychological disorders. Among these three ratings, there was substantial — but imperfect — agreement, with a Krippendorf's alpha of .722. It is possible that, in the case of depression, a "gold-standard" label simply does not exist.

Lay understandings of depression diverge even further from psychiatric understandings of depression (Davidsen and Fosgerau, 2014). Rather than focusing solely on established criterion, many individuals with depression use vivid metaphors that richly convey the lived experience of depression (Karp, 2016). We see this in our data as well. Participants use metaphors such as *a smoking gun of sadness* and *a rug pulled out from under me.* Another searches aloud to find a good metaphor in the interview, *a bird in a cage, a fish that cant swim in water, a bird without wings.*

We also see in our data how participants carefully — but inconsistently — distinguish between depression, happiness, contentment, and other states and moods. For example, when asked, *when was the last time you felt really happy?'* one participant clarifies, *what type of happiness are you looking for?* Another participant mentions being a determined individual and says, despite having some *deep depression*, *i work myself into being in positive states of mind.* Meanwhile, another says, *i i i don't know if this sounds right but i'm not seeking happiness i i can only explain it as i'm content* [1]. Others echo the desire for contentment over happiness,

happiness is a is a true and permanent state of mind I think I'm far more interested in it in contentment I'm far more interested in it purpose and in that yeah contentment person purpose and and a sense of metal metal [sic] involvement engagement and also [HESITATION] I guess and it is that would be my happiness. [2]

	Krippendorf's alpha *N=1523 utterances Rated three times each*
Depression	.18
Lack of Interest	.078
Depressed Mood	.19
Sleep	.16
Low Energy	.15
Appetite	.062
Low Self-Esteem	.14
Trouble Concentrating	.065
Psychomotor Impairment	.059
Symptoms from the Patient Health Questionnaire-8.	

Table 1: Crowd workers agreement on depression symptoms

In the above excerpt, as in many others, we see how participants may make meaning of their experiences and feelings as they try put them into words out loud. In the above case, for example, the participant initially reports being more interested in contentment and purpose than happiness, proceeds to describe contentment, and then returns to equating this description of contentment and purpose to their happiness.

In lay annotations of depression and depression symptoms, we also find a lack of agreement on depression. Specifically, we found that crowd workers had only slight agreement on the whether a speaker might have depression or depression symptoms, based on excerpts of transcribed interview data. For agreement on whether the speaker might have depression or not (or if there is no evidence from the utterance) agreement was slight (Krippendorf's alpha = .18). Among agreements on specific symptoms, the average Krippendorf's alpha across all PHQ-8 symptoms was .12, suggesting little or no agreement on text representing these symptoms. This further varied by symptom, as can be seen in table 1.

A total of 381 (25%) of utterances had perfect agreement on depression, when agreement was measured as unlikely (or very unlikely), no evidence, or likely (or very likely). Among these, 146 (38%) were agreements on very unlikely or unlikely, 224 (59%) were agreements on likely or very likely, and 11 (3%) were agreements on the lack of evidence. It is possible that certain types of evidence are easier to detect than other

[1] Repetitions are common parts of human conversational language

[2] Note this is how the original verbal data was machine-transcribed.

	Spearman Correlation *N=1523 utterances Median of three ratings*
Depression	.29
Lack of Interest	.20
Depressed Mood	.21
Sleep	.12
Low Energy	.15
Appetite	.11
Low Self-Esteem	.21
Trouble Concentrating	.10
Psychomotor Impairment	.061
Symptoms from the Patient Health Questionnaire-8. All correlations are significant at $p < .01$.	

Table 2: Correlation between crowd workers ratings of depression symptoms in utterances and PHQ-8 scores of speakers

types of evidence, especially evidence *for* mental distress. Overall, crowd workers' ratings were weakly associated with speakers' symptoms according to the PHQ-8. Higher median ratings of depression tended to be associated with slightly higher scores on the PHQ-8 (Spearman ρ= .29, $p < .01$). Associations strengths varied further by symptom, as shown in table 2. These low rates of interrater agreement (and low correlation between PHQ scores and lay annotations) may not be surprising. Emotional states and moods are notoriously difficult to annotate, particularly attempts to annotate emotions beyond basic ones such as anger, joy, and sadness (Devillers et al., 2005). Depression is further complicated in that it is not merely constituted by feelings but also somatic and cognitive impairment.

Interestingly, we do find evidence that perceptions of depression may be related to known features such as the use of pronouns and talk of sadness. In particular, we find that among utterances with perfect agreement, utterances are more likely to be rated for depression if they contain more first-person singular ($p = .01$), less first-person plural ($p = .001$), contain more talk of negative emotions ($p < .001$) and, in particular, sadness ($p < .001$), and less talk of positive emotions ($p < .001$). We find other intuitive patterns as well, such as that utterances with more talk of health ($p < .001$), and less talk of leisure($p < .001$), tend to be rated as depressed more often than not depressed.

4.2 Beyond the binary: mental health as a spectrum of symptoms

Most of nature is continuous and dimensional, and psychological distress is no exception (Bowins, 2015; Insel et al., 2010; Adam, 2013; Kapur et al., 2012; Andrews et al., 2007; Lewinsohn et al., 2000; Nelson et al., 2017). However, humans tend to categorize the continuous; such as labeling an individual as depressed or not. Categorization enables us to more rapidly process information, but also blurs the intricacies of a phenomena. Mental health categories can also validate the illness experience, improve diagnostic reliability, provide some common language (e.g., for medical billing), and suggest clues for treatments. However, diagnostic thresholds for depression hold limited clinical significance and even sub-threshold symptoms are associated with a decline in well-being (Lewinsohn et al., 2000). And so, for all our carefully constructed categories, we must move past a categorical approach to mental illness (Insel et al., 2010; Adam, 2013; Kapur et al., 2012; Jackson et al., 2017; Lewinsohn et al., 2000).

Luckily, computational models do not need the same heuristics that we need to efficiently process information. These models can capture depression (or mental illness at large) more realistically — as something continuous, dimensional, and multifaceted. The majority of the published models reviewed in this paper examine depression as a binary phenomenon. At the least, models should detect depression as a continuous phenomenon, such as PHQ-8 score.

Performance metrics and visuals based on categorical conceptions of depression (such as sensitivity) are still useful for human readers. But the underlying model should model depression as continuous. Ideally, we would consider depression in more dimensions, such as duration of depression episode, depression history, and the amount of impairment caused by the episode (Bowins, 2015; Andrews et al., 2007). Indeed, literature already suggests that, like the cognitive impairment associated with depression, linguistic patterns vary by duration of depression episode (Capecelatro et al., 2013). Furthermore, Tsugawa et al. find that depression of social media users is best predicted by a window of two months of social media expression, rather than a larger or smaller window of time.

It may also be fruitful to detect *symptoms* of de-

pression, rather than aiming to detect depression itself. In fact, some scholars reject the notion that depression exists as a latent entity causing observable symptoms — also known as the latent-disease model. Instead, what we consider depression is a causal, mutually reinforcing chain of symptoms (Nelson et al., 2017; Wichers et al., 2016; Wichers, 2014; Borsboom and Cramer, 2013; van Borkulo et al., 2015). In other words, depression is a dynamic system stuck in feedback loops. These scholars suggest depression should be studied with relevant, cross-disciplinary tools and theories, such as dynamical systems theory to consider tipping points and phase transitions in the depression experience, and network theory to model depression as a network of symptoms.

A symptom-based approach would also account for diversity of symptoms that may constitute distress. This might provide another approach address recent concerns about the external validity of depression models to culture and gender compositions of populations (De Choudhury et al., 2016; Tsugawa et al., 2015). Research using clinical texts, namely medical notes, has already begun to move in a symptom-based direction with success and may provide inspiration (Jackson et al., 2017). Whether we detect symptoms or overall depression score, it is important to consider that some symptoms of depression (e.g., somatic symptoms) might be more or less prevalent in language compared to their morbidity, and stronger or weaker predictors of distress when present.

A symptom-based and continuous approach to modeling could also help us move towards modeling how depression overlaps many symptoms of post-traumatic stress disorder, anxiety, and other mental illnesses. Indeed, mental illnesses 1) are often co-morbid, 2) share many of the same symptoms and 3) may exacerbate each other (Kessler et al., 1994). In fact, general practitioners often informally regard concomitant symptoms of mental distress (such as symptoms of an eating disorder, depression, and anxiety) as manifestations of one underlying condition of mental distress rather than symptoms of multiple distinct conditions (Davidsen and Fosgerau, 2014). They use diagnostic tools primarily due to pressure from psychiatric medicine and for insurance purposes (Davidsen and Fosgerau, 2014). In our data, participants also often talk about multiple mental illnesses at once, and discussion of symptoms may not be clearly attributed to one condition or another. One participant, for example, talks about *the anxiety part of my depression* as if they are one of the same. Another participant suggests that their depression is even caused by anxiety, saying *eh eh just so many things i worry about and that's what was making me depressed.* Another reflects, *depression kind of goes with anxiety if it's not under control.* Thus, a more holistic approach to detecting mental health might enable greater sensitivity to different expressions of mental distress rather than fixating on categories of "depression" which were constructed by psychiatric medicine.

In our data, we also find preliminary evidence that linguistic patterns vary by symptom, not just depression severity. We investigated how known linguistic markers of depression based on LIWC, such as the use of negative emotions, vary by depression symptom. As mentioned earlier, we measure eight symptoms based on the PHQ.

We illustrate a few of these results in figure 1, to show the use of sadness words for each of the eight PHQ-8 symptoms, as well as for binary measures of depression based on aggregating these symptoms (for reference). As expected from previous research, those categorized as depressed tend to use more words about sadness than those not categorized as depressed ($P < .001$). This pattern, however, appears exaggerated when we look at talk of sadness among those who report more severe levels of depressed mood, versus milder levels ($P < .001$). Indeed, those reporting high levels of depressed mood use more words about sadness than do those reporting high levels of depression ($P = .03$) [3].

Perhaps specific symptoms of depression, such as depressed mood, could be driving the relationship between depression and certain lexicon. If so, predictors based on this lexicon could systematically miss individuals who express depression more in terms of symptoms such as a lack of interest — the use of words about sadness does not seem to differ by someone's lack of interest or ability to take pleasure in their experiences ($P = .68$). More broadly, it is possible that certain linguistic markers are better predictors of certain symptoms than others. Thus errors from models predicting depression should be carefully investigated for patterns in errors. Perhaps models de-

[3] Statistical comparisons between groups reporting high severity levels should be interpreted with caution, as these are not independent groups.

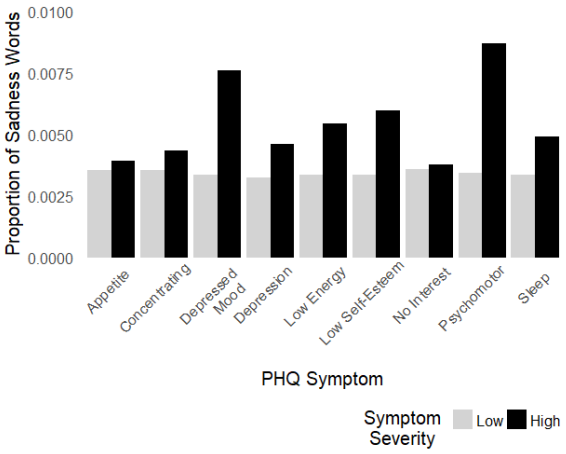

Figure 1: Proportion of "Sadness" words by PHQ-8 Symptom

tecting mental distress might also make more informative predictions about depression *symptoms* rather than depression overall. A symptom based approach would have the added benefit of more realistically portraying the facets of depression and being more generalizable across different expressions of symptoms.

5 Measuring model performance

A great deal of research goes into assessing the performance of predictive models. There are F1 scores, accuracy rates, recall, ROC curves, precision, and root mean-square error, among other measures (Steyerberg et al., 2010).

In most mental health contexts, the most costly error is to miss an individual with depression. Thus, models should prioritize capturing depression among those who have depression. A performance metric commonly used with this in mind is sensitivity, also called recall [4]. Specificity and precision rates, on the other hand, may be useful even if somewhat low. Specificity refers to the proportion of those without depression who are correctly detected as not having depression. Precision rates refer to the proportion of those actually with depression out of all those classified as having depression. Even low rates for specificity are useful to "weed out" a chunk of individuals not at risk. Particularly if a tool to detect mental health is used a screening tool in a clinical setting, this reduces the burden of more extensive screens and doctor evaluations. While most studies reviewed in this paper do not explicitly discuss of which measures they prioritize, one study stands out in that the metrics of candidate models are reflexively considered, based on deployment goals of models (Nadeem, 2016). The authors prioritize recall over sensitivity, and accuracy over F1-score, when comparing candidate models.

It may be fruitful to compare clinicians' diagnostic practices with computational models. For example, Resnik et al. compared computational predictions of depression with predictions made by three practicing clinical psychologists. They used binary measures of depression from the Beck Depression Inventory (BDI), with a standard cut-off of 14. The psychologists' sensitivities to the BDI (.83, .83, and .66 respectively) were far higher than the models (average of .50), while their precision was far lower than models (.38, .33, and .33, respectively among raters, and average of .47 among models). Perhaps part of this sensitivity is humans' tendency to heavily weigh evidence *for* depression over any other information - including evidence against depression. In developing our models, we also need to account for this trade-off. Like humans detecting mental health, in building automated methods to detect depression we may need to be willing to work with low specificity and precision to enable with greater sensitivity.

In considering performance metrics, we can gain insight from disciplinary standards in medicine to release new diagnostics screening tools, such as the PHQ-8. For example, unlike publications of diagnostic screening tools in medicine, many studies reviewed in this paper do not present confidence measures on performance. Further, as also noted by Guntuku et al., an issue with sensitivity is that it depends on the prevalence of a condition. Thus sensitivities of a model are difficult to compare across datasets. In medicine, another commonly used performance metric which addresses this issue is positive predictive value. And, like practices in medicine, modeling efforts might consider using a single model across various populations to understand how it generalizes to new, unique groups of people.

6 Conclusions

A flurry of recent research has produced various models for detecting depression and other mental health outcomes. As exciting as the prospects of

[4] Sensitivity, or recall, here is the proportion of those with depression who are correctly detected as having depression.

such tools are, they also stir up old debates and new on the computational representation of mental health.

Most importantly, this paper urges the careful consideration of labels in models of mental health. At the least, depression should be modeled a continuous rather than binary outcome, and models might detect specific symptoms in addition to detecting depression as an overall construct. A reconsideration of labels in the field of modeling mental health is timely. Recently, the NIMH has also drawn attention to weakness of current classifications of mental health. The NIMH is now working to transform psychiatric diagnoses to acknowledge the dimensionality of mental health (Insel et al., 2010). Meanwhile, a growing movement in psychiatry calls for a re-acquaintance with phenomenology. Categories for mental health risk being so articulated and abstracted that they lose touch with the diversity of illness experiences (Andreasen, 2006; Jacob, 2012; Mullen, 2006).

Given the diversity in how mental distress is expressed, and lack of a gold standard, model performance and errors should be evaluated in depth. For example, there might be consistent types of symptoms, or depression experiences, not being detected. And, it is possible that certain linguistic features may be better predictors of certain symptoms (or types of depression experiences) than others.

Meanwhile, while presenting and comparing model performances, we need to be careful about compounding inaccuracies. Even if a model is published with quantifications of modeling error, these quantifications do not include error at capturing depression - only the proxy used to capture depression, such as the PHQ. If the PHQ and other self-report measures are imperfect, and we use these as a gold standards without acknowledging their limitations, this inflates the true error rate of our models.

While the search for valid constructs of mental health is still underway, an ideal data-set would include multiple physicians ratings as well as a variety of other clinical and non-clinical measures of depression. In turn, comparing errors across these metrics might also shed light on the nature of mental distress itself.

While research in this area has recently focused on the production of high-performing models, it seems likely that literature will soon reach saturation in the number of published models. Now, models will need to be reflexively tuned, borrowing additional insight from areas such as medicine and social sciences. Modeling goals might now also include feasibility of deployment and generalizability.

It may help to a step back to move forwards. Most importantly, we need to reconsider our understanding of mental illness and be precise about what, in fact, we are detecting. And we need to consider how to develop predictive models that incorporate the uncertainty in our understanding of depression and other cultural idioms of distress. Research efforts can then turn to realizing the vision that initially motivated these models: their deployment for early, scalable, and low-burden intervention and diagnosis of depression.

Acknowledgments

The authors thank the members of the Health Working Group at UCLA and the anonymous reviewers for their feedback. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1650604. This material is also based upon work supported by the U.S. Army Research Laboratory and DARPA under contract number W911NF-14-D-0005. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Government, and no official endorsement should be inferred.

References

Ahmed Aboraya, Eric Rankin, Cheryl France, Ahmed El-Missiry, and Collin John. 2006. The reliability of psychiatric diagnosis revisited: The clinician's guide to improve the reliability of psychiatric diagnosis. *Psychiatry (Edgmont)*, 3(1):41.

David Adam. 2013. Mental health: On the spectrum. Accessed: 2018-03-23.

Nancy C Andreasen. 2006. Dsm and the death of phenomenology in america: an example of unintended consequences. *Schizophrenia bulletin*, 33(1):108–112.

Gavin Andrews, Traolach Brugha, Michael E Thase, Farifteh Firoozmand Duffy, Paola Rucci, and Timothy Slade. 2007. Dimensionality and the category of major depressive episode. *International Journal of Methods in Psychiatric Research*, 16(S1).

Bruce Arroll, Felicity Goodyear-Smith, Susan Crengle, Jane Gunn, Ngaire Kerse, Tana Fishman, Karen Falloon, and Simon Hatcher. 2010. Validation of phq-2 and phq-9 to screen for major depression in the primary care population. *The Annals of Family Medicine*, 8(4):348–353.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Herbert Blumer. 1986. *Symbolic interactionism: Perspective and method*. Univ of California Press.

Claudia van Borkulo, Lynn Boschloo, Denny Borsboom, Brenda WJH Penninx, Lourens J Waldorp, and Robert A Schoevers. 2015. Association of symptom network structure with the course of depression. *JAMA psychiatry*, 72(12):1219–1226.

Denny Borsboom and Angélique OJ Cramer. 2013. Network analysis: an integrative approach to the structure of psychopathology. *Annual review of clinical psychology*, 9:91–121.

Brad Bowins. 2015. Depression: discrete or continuous? *Psychopathology*, 48(2):69–78.

Philip Burnard. 1991. A method of analysing interview transcripts in qualitative research. *Nurse education today*, 11(6):461–466.

Rafael A Calvo, David N Milne, M Sazzad Hussain, and Helen Christensen. 2017. Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*, 23(5):649–685.

Glorisa Canino and Margarita Alegría. 2008. Psychiatric diagnosis–is it universal or relative to culture? *Journal of Child Psychology and Psychiatry*, 49(3):237–250.

Maria R Capecelatro, Matthew D Sacchet, Peter F Hitchcock, Samuel M Miller, and Willoughby B Britton. 2013. Major depression duration reduces appetitive word use: An elaborated verbal recall of emotional photographs. *Journal of psychiatric research*, 47(6):809–815.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60.

Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F Quatieri. 2015. A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71:10–49.

Annette S Davidsen and Christina F Fosgerau. 2014. What is depression? psychiatrists and gps experiences of diagnosis and the diagnostic process. *International journal of qualitative studies on health and well-being*, 9(1):24866.

Kenneth Davison. 2006. Historical aspects of mood disorders. *Psychiatry*, 5(4):115–118.

Munmun De Choudhury, Scott Counts, Eric J Horvitz, and Aaron Hoff. 2014. Characterizing and predicting postpartum depression from shared facebook data. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 626–638. ACM.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. *ICWSM*, 13:1–10.

Munmun De Choudhury, Sanket Sharma, Tomaz Logar, Wouter Eekhout, René Nielsen, Georgia Tech, and Global Pulse. 2016. Quantifying and understanding gender and cross-cultural diferences in mental health expression via social media.

David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, et al. 2014. Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 1061–1068. International Foundation for Autonomous Agents and Multiagent Systems.

Laurence Devillers, Laurence Vidrascu, and Lori Lamel. 2005. Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, 18(4):407–422.

Kathleen C Fraser, Frank Rudzicz, and Graeme Hirst. 2016. Detecting late-life depression in alzheimer's disease through analysis of speech and language. In *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*, pages 1–11.

Simon Gilbody, David Richards, Stephen Brealey, and Catherine Hewitt. 2007. Screening for depression in medical settings with the patient health questionnaire (phq): a diagnostic meta-analysis. *Journal of general internal medicine*, 22(11):1596–1602.

Jonathan Gratch, Ron Artstein, Gale M Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, et al. 2014. The distress analysis interview corpus of human and computer interviews. In *LREC*, pages 3123–3128. Citeseer.

Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49.

Rahul Gupta, Nikolaos Malandrakis, Bo Xiao, Tanaya Guha, Maarten Van Segbroeck, Matthew Black, Alexandros Potamianos, and Shrikanth Narayanan.

2014. Multimodal prediction of affective dimensions and depression in human-computer interactions. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 33–40. ACM.

Uriel Halbreich and Sandhya Karkun. 2006. Cross-cultural and social diversity of prevalence of postpartum depression and depressive symptoms. *Journal of affective disorders*, 91(2):97–111.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Christine Howes, Matthew Purver, and Rose McCabe. 2014. Linguistic indicators of severity and progress in online text-based therapy for depression. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 7–16.

Frederick Y Huang, Henry Chung, Kurt Kroenke, Kevin L Delucchi, and Robert L Spitzer. 2006. Using the patient health questionnaire-9 to measure depression among racially and ethnically diverse primary care patients. *Journal of general internal medicine*, 21(6):547–552.

Thomas Insel. 2013. Transforming diagnosis. Accessed: 2018-03-23.

Thomas Insel, Bruce Cuthbert, Marjorie Garvey, Robert Heinssen, Daniel S Pine, Kevin Quinn, Charles Sanislow, and Philip Wang. 2010. Research domain criteria (rdoc): toward a new classification framework for research on mental disorders.

Richard G Jackson, Rashmi Patel, Nishamali Jayatilleke, Anna Kolliakou, Michael Ball, Genevieve Gorrell, Angus Roberts, Richard J Dobson, and Robert Stewart. 2017. Natural language processing to extract symptoms of severe mental illness from clinical text: the clinical record interactive search comprehensive data extraction (cris-code) project. *BMJ open*, 7(1):e012012.

KS Jacob. 2012. Psychiatric assessment and the art and science of clinical medicine. *Indian journal of psychiatry*, 54(2):184.

Shitij Kapur, Anthony G Phillips, and Thomas R Insel. 2012. Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? *Molecular psychiatry*, 17(12):1174.

David A Karp. 2016. *Speaking of sadness: Depression, disconnection, and the meanings of illness*. Oxford University Press.

Ronald C Kessler and Evelyn J Bromet. 2013. The epidemiology of depression across cultures. *Annual review of public health*, 34:119–138.

Ronald C Kessler, Katherine A McGonagle, Shanyang Zhao, Christopher B Nelson, Michael Hughes, Suzann Eshleman, Hans-Ulrich Wittchen, and Kenneth S Kendler. 1994. Lifetime and 12-month prevalence of dsm-iii-r psychiatric disorders in the united states: results from the national comorbidity survey. *Archives of general psychiatry*, 51(1):8–19.

Kurt Kroenke and Robert L Spitzer. 2002. The phq-9: a new depression diagnostic and severity measure. *Psychiatric annals*, 32(9):509–515.

Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. 2001. The phq-9. *Journal of general internal medicine*, 16(9):606–613.

Peter M Lewinsohn, Ari Solomon, John R Seeley, and Antonette Zeiss. 2000. Clinical implications of" subthreshold" depressive symptoms. *Journal of abnormal psychology*, 109(2):345.

Paul E Mullen. 2006. A modest proposal for another phenomenological approach to psychopathology. *Schizophrenia bulletin*, 33(1):113–121.

Moin Nadeem. 2016. Identifying depression on twitter. *arXiv preprint arXiv:1607.07384*.

Barnaby Nelson, Patrick D McGorry, Marieke Wichers, Johanna TW Wigman, and Jessica A Hartmann. 2017. Moving from static to dynamic models of the onset of mental disorder: a review. *JAMA psychiatry*, 74(5):528–534.

Thin Nguyen, Dinh Phung, Bo Dao, Svetha Venkatesh, and Michael Berk. 2014. Affective and content analysis of online depression communities. *IEEE Transactions on Affective Computing*, 5(3):217–226.

Gordon Parker, Y-C Cheah, and K Roy. 2001. Do the chinese somatize depression? a cross-cultural study. *Social psychiatry and psychiatric epidemiology*, 36(6):287–293.

Eugene S Paykel. 2008. Basic concepts of depression. *Dialogues in clinical neuroscience*, 10(3):279.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015. Technical report.

Jennifer Radden. 2003. Is this dame melancholy?: Equating today's depression and past melancholia. *Philosophy, Psychiatry, & Psychology*, 10(1):37–52.

Andrew G Reece, Andrew J Reagan, Katharina LM Lix, Peter Sheridan Dodds, Christopher M Danforth, and Ellen J Langer. 2017. Forecasting the onset and course of mental illness with twitter data. *Scientific reports*, 7(1):13006.

Philip Resnik, Anderson Garron, and Rebecca Resnik. 2013. Using topic modeling to improve prediction of neuroticism and depression in college students. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1348–1353.

Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. 2004. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8):1121–1133.

H Andrew Schwartz, Johannes Eichstaedt, Margaret L Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. 2014. Towards assessing changes in degree of depression through facebook. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 118–125.

Galit Shmueli. 2010. To explain or to predict? *Statistical science*, pages 289–310.

Ewout W Steyerberg, Andrew J Vickers, Nancy R Cook, Thomas Gerds, Mithat Gonen, Nancy Obuchowski, Michael J Pencina, and Michael W Kattan. 2010. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.)*, 21(1):128.

Shin Tarumi, Atsushi Ichimiya, Shin Yamada, Masahiro Umesue, and Toshihide Kuroki. 2004. Taijin kyofusho in university students: patterns of fear and predispositions to the offensive variant. *Transcultural psychiatry*, 41(4):533–546.

Brett D Thombs, Andrea Benedetti, Lorie A Kloda, Brooke Levis, Ioana Nicolau, Pim Cuijpers, Simon Gilbody, John PA Ioannidis, Dean McMillan, Scott B Patten, et al. 2014. The diagnostic accuracy of the patient health questionnaire-2 (phq-2), patient health questionnaire-8 (phq-8), and patient health questionnaire-9 (phq-9) for detecting major depression: protocol for a systematic review and individual patient data meta-analyses. *Systematic reviews*, 3(1):124.

Sho Tsugawa, Yusuke Kikuchi, Fumio Kishino, Kosuke Nakajima, Yuichi Itoh, and Hiroyuki Ohsaki. 2015. Recognizing depression from twitter activity. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3187–3196. ACM.

M Wichers. 2014. The dynamic nature of depression: a new micro-level perspective of mental disorder that meets current challenges. *Psychological medicine*, 44(7):1349–1360.

Marieke Wichers, Peter C Groot, ESM Psychosystems, EWS Group, et al. 2016. Critical slowing down as a personalized early warning signal for depression. *Psychotherapy and psychosomatics*, 85(2):114–116.

James R Williamson, Elizabeth Godoy, Miriam Cha, Adrianne Schwarzentruber, Pooya Khorrami, Youngjune Gwon, Hsiang-Tsung Kung, Charlie Dagli, and Thomas F Quatieri. 2016. Detecting depression using vocal, facial and semantic communication cues. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 11–18. ACM.