A method for identifying predictive markers of mental illness in social media data

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

Citable link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:40046396

Terms of Use
This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA
A method for identifying predictive markers of mental illness in social media data

A dissertation presented
by
Andrew Garrett Reece
to
The Psychology Department
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Psychology

Harvard University
Cambridge, Massachusetts

November 2016
A method for identifying predictive markers of mental illness on social media

Abstract

Undiagnosed mental illness poses a significant health risk. In-person screenings to identify individuals at-risk of mental illness are expensive, time-consuming, and often inaccurate. This report presents an array of computational methods that can be used to identify predictive markers of mental illness - specifically, depression and Post-Traumatic Stress Disorder - by scanning and interpreting text and images posted on social media. Separate analyses of Twitter and Instagram data are presented. Predictive features were extracted from social media posts (NTwitter = 279,951, NInstagram = 43,950) using a variety of techniques, including color analysis, face detection, semantic analysis, and Natural Language Processing. Resulting models successfully discriminated between depressed and healthy content, and compared favorably to general practitioners’ average success rates in diagnosing depression. Results held even when the analysis was restricted to content posted before first depression diagnosis. In the case of Twitter data, state-space temporal analysis suggests that onset of depression may be detectable from Twitter data several months prior to diagnosis. These methods offer a data-driven, predictive approach for early screening and detection of mental illness.
# Table of Contents

Abstract.................................................................................................................. iii
Table of Contents.................................................................................................... iv
Dedication............................................................................................................... v
Acknowledgments................................................................................................. vi
Chapter I: Introduction............................................................................................ 1
Chapter II: Overview of Methods........................................................................... 8
Chapter III: Depression and Instagram................................................................. 29
  Methods.............................................................................................................. 34
  Results................................................................................................................. 47
  Discussion........................................................................................................... 60
Chapter IV: Forecasting mental illness on Twitter............................................... 66
  Methods.............................................................................................................. 70
  Results................................................................................................................. 81
  Discussion........................................................................................................... 90
Chapter V: Concluding thoughts............................................................................ 96
References............................................................................................................. 112
Appendix................................................................................................................ 125
For my father, Gary
Acknowledgments

I thank my advisor, Dr. Ellen Langer, for her gracious mentorship and invaluable tutelage.

I thank my parents for their unflagging support and encouragement.

At the time of this writing, modified and condensed versions of Chapters III and IV have been submitted for publication as jointly-authored works. I thank my co-authors, A. Reagan, C. Danforth, E. Langer, K. Lix, and P. Dodds, for their invaluable contributions to our work; it was a pleasure to conduct science with you all. Any errors or omissions in the versions of our work reported here are entirely my own.

I thank D. Gergishanova, my research assistant, for her help in recruitment and data collection.

I thank D. Yudkin and G. Cooney for valuable discussions on how to do good and interesting science.
Chapter I: Introduction

Thanks to the mass popularization of social media, vast amounts of psychologically relevant data are created every day. By the time you finish reading this sentence, 20,000 new tweets will have appeared on Twitter (InternetLiveStats, 2016). Over 100 million new photos and videos are shared daily on Instagram, and Facebook sees well over 30 billion pieces of content shared each month (Instagram, 2016; InternetLiveStats, 2016). The thoughts, feelings, musings, interests, and social relationships of billions of people are neatly stored and categorized by these services, and much of it is openly available to the public. Despite the abundance of this social data, the powerful information encoded in it has remained vastly underutilized by psychologists. The emerging field of computational social science is just beginning to harness psychological features in data for research purposes.

One approach is to use past information to better understand how people behave in the present – or how they will behave in the future. On a basic level, social media records are akin to diaries, albeit extremely well-documented, usually public ones. They are the collated, time-stamped, metadata-enhanced records of thoughts and feelings people shared in the moment. For that matter, social media records store not only the content of a user’s postings, but also patterns of activity – posting frequency, average post length, average number of replies to other users, and so on. These patterns can act as proxies for social engagement, and motivation in general. The benefit of having these measures marked clearly in time is that it becomes possible to track psychological change.

This presents a golden opportunity to clinical psychologists, who frequently have to rely on patients’ imperfect recollections of symptoms and experiences, in order to make diagnoses. With the ubiquity of social media in many people’s lives today, it’s not hard to imagine that a
trail of psychological breadcrumbs is laid out in the tweets, pokes, likes, and status updates that make up the daily online activity of the modern social media enthusiast. The key, then, to harnessing this trove of information is to develop a good breadcrumb detector – a method for identifying, extracting, and analyzing social media archives that can successfully reveal markers of psychological change.

Early prototypes have already emerged in the field of computational social science. For example, Dodds et al. (2011) showed that the happiness of language used on Twitter demonstrates temporal periodicity, rising and falling predictably based on the day of the week (Saturdays are happy, Tuesdays are sad) and time of day (people are happiest after they leave work). Correa et al. (2010) found reliable measures of “Big Five” personality traits in the usage patterns of Twitter users. To date, however, the most popular topic in this developing literature has been concerned with identifying markers of mental and physical illnesses in social media data.

De Choudhury, Gamon, Counts, & Horvitz (2013) found that Twitter users’ posts in the weeks leading up to a major depressive episode differed significantly on a number of measures from the tweets of a comparison group. De Choudhury, Counts, Horvitz, & Hoff (2014) replicated these results with Facebook user data. Katikalapudi, Chellappan, Montgomery, Wunsch, & Lutzen (2012) found similar depression trends with undergraduate Twitter users at a Midwestern US university. A study of female Twitter users who had recently given birth showed significant changes in postpartum Twitter activity and language use, compared to a control cohort (De Choudhury, Counts, & Horvitz, 2013a; De Choudhury, Counts, & Horvitz, 2013b). Coppersmith, Harman, & Dredze (2014) employed a similar methodology to De Choudhury et al. (2013) to demonstrate that individuals suffering from Post-Traumatic Stress
Disorder (PTSD) used different patterns of language in Twitter postings. Notably, many of these studies are authored by the same groups of researchers, which underscores the fact that this field is still in its infancy.

In this report, I present a series of novel methodological contributions for the identification and tracking of mental illness - particularly depression and PTSD - using data culled from Twitter and Instagram. The Twitter findings I describe correct a number of methodological limitations of the previous work cited above, and add a new method for tracking the emergence of mental illness over time. The Instagram study demonstrates how computational image processing methods may be used to extract psychologically-meaningful information from visual social media, such as photographs posted to Instagram. Together, the studies presented in this report represent not only novel methodological contributions to the field of psychological science, but also provide new potential for both applied and theoretical advances. Chapter II discusses a number of methodological considerations which are common to both my Instagram and Twitter research. Chapter III is the report of my Instagram findings, and Chapter IV is the report of my Twitter findings. Chapters III and IV are meant to read as stand-alone research reports. In Chapter V, I conclude with an address of limitations and a review of possible avenues for future research.
Chapter II. Overview of methods

The methods used in this research are concerned with finding patterns in social media use that reliably segregate healthy and affected users. Specific methods are described in detail in Chapters III and IV; in this chapter, I present the step-by-step methodological outline to which each study adheres, along with a treatment of the various problems and solutions associated with each step.

Step 1: Identification of platform-specific data

Most social media platforms provide an Application Programming Interface (API) for developers, which allows for the querying and retrieval of posts and metadata (see dev.twitter.com and www.instagram.com/developer). I chose to study Instagram and Twitter. This choice was primarily motivated by convenience; their APIs are easier to use, and offer better access to user data, than other platforms (e.g., Facebook). A secondary motivating factor was that Twitter and Instagram are noted for their use of text and images, respectively, as primary modes of communication, and I wanted to show that the analysis for psychological markers of mental illness could be performed on both text and image data.

Determining which metadata fields might yield useful psychological information was a principal task in this project. A single photograph posted to Instagram, for example, carries several dozen fields of associated metadata. These fields provide information about post timing and location, user-generated captions and hashtags, poster account details, and community reactions (e.g., likes, comments). Many of these metadata can be further parsed to extract

1 I use the term “affected” to refer to the sample data collected from participants who were diagnosed with a mental health condition (specifics differ per study). The term “healthy” refers to sample data collected from participants who reported no history of that mental health condition, although the label is something of a misnomer, as healthy
additional features. For example, the comments associated with a tweet might be analyzed for the frequency of a specific word, like “sad” or “happy”. Geotagged posts provide GPS coordinates which can be used to generate location features (e.g. “tends to tweet from coffee shops”, or “likely city of residence”). As the studies conducted all employed observational designs, hypothesis testing focused on the quality of selected predictive features and their capacity to distinguish between healthy and affected users. The introductory sections of Chapters III and IV explore the rationale behind selected features in detail.

**Step 2: Recruitment and data collection**

Finding people who are willing to share both personal mental health histories and social media data for the sake of academic research is a substantial challenge. While a host of technical challenges also presented themselves in this research, the difficulties posed by a sparsity of recruitment channels are paramount, as without usable data, there can be no analysis. As such I devote more extensive consideration here to the current data collection strategies used in computational social science, highlighting the strengths and weaknesses most salient to my own research. Most social media analytics research has used one of two approaches to recruitment and data collection, which I refer to ask *bulk ingestion* and *active recruitment*. I review each method, focusing on previous studies which have used Twitter to study depression or PTSD\(^2\).

*Bulk Ingestion*

Bulk ingestion bypasses participant recruitment entirely, and relies instead on the premise that some social media platforms make content publicly available as part of their terms of use. The Twitter API allows collection of the most recent 3,200 tweets publicly posted by any account, given that that account username is known by the API user (Twitter, 2016). Twitter

\(^2\) To date, almost no studies have been conducted on Instagram; those that did used variants of bulk ingestion for data collection (Andalibi, Ozturk, & Forte, 2015; Lup, Trub, & Rosenthal, 2015).
also allows for streaming data collection, for which a small percentage of all public tweets (widely believed to be around 1-2%, although Twitter keeps the exact number secret) can be ingested in real-time\(^3\). While Twitter and other platforms impose rate limits which restrict the volume of data that can be collected in a given time span, over time it is relatively easy to amass a large corpus of social media posts, without ever having interacted with any of the individuals who authored those posts.

The main benefit of the bulk ingestion approach is volume. Hundreds of millions of data points can be collected in a few weeks. The main drawback is a lack of specificity. Without any direct interaction with individuals represented in collected data, it is hard to know which posts are interesting for a given topic of study. Alternate methods of identifying data points of interest become necessary. In the context of studying specific health conditions, the most common approach has been to search all the collected tweets for explicit indications of diagnosis disclosures, e.g. tweets containing the text “Today I was diagnosed with depression”. A second round of data collection can then zero in on accounts which turned up in such an “explicit disclosure” search, and the 3,200 most recent tweets posted by those accounts can be stored for further analysis. A comparison sample of “healthy” Twitter users is usually identified by choosing usernames at random from the initial bulk ingestion, and repeating the same process of acquiring those users’ 3,200 most recent tweets.

The bulk ingestion approach can be useful for a first-pass analysis, but is hampered by a number of data quality issues which prevents it from more targeted analysis. Let us assume, for the moment, that when a Twitter account posts a tweet reading, “I was diagnosed with

\(^3\) Note that, with streaming collection, the random availability of tweets means that collecting a single tweet from a given user does not mean that any other tweets by that same user will appear in the stream.
depression”, it means that the account holder was actually diagnosed with depression⁴. Unless a specific temporal reference is included with a disclosure (e.g. “Today I was diagnosed with depression”), it is impossible to know which posts in a user’s collected history were posted before or after receiving a diagnosis (or, for that matter, beginning to exhibit depressive symptoms). Any analysis based on bulk ingestion may include both post-diagnosis posts, and post-treatment/post-recovery posts.

Post-diagnosis posts present a potential confound when trying to determine whether a mental health condition itself is responsible for a change in communication patterns. Previous research has shown that individuals diagnosed with depression come to self-identify with their diagnosis, and begin to express themselves with this identity in mind (Cornford, Hill, & Reilly, 2007; Karp, 1994). Accordingly, predictive markers of depression observed in post-diagnosis tweets, for example, might be caused by actual signals of depressive symptoms, or might instead be indicative of a conscious effort to represent oneself as depressed.

A second category of limitations with the bulk ingestion approach concerns inter-group crossover between depressed and healthy samples, a form of data contamination. Recall that healthy samples are established by randomly selecting usernames from a large sample of tweets. No effort is made to screen out users who might actually be affected by the health condition of interest, and the assumption of health relies on sheer volume, rather than on quality control. For example, while Coppersmith et al. (2014) identified just 244 Twitter users who explicitly disclosed PTSD diagnoses in their tweets, the authors compiled a comparison sample of 5,728 users (all using bulk ingestion). As PTSD occurs in a very small proportion of the general population (Stein et al., 2000), and as there is no compelling argument to be made that the

⁴ While this is not an entirely justified assumption, this approximates the trust extended to in-person study participants’ self-reporting of psychological conditions, and so is a broader methodological concern in behavioral studies, not limited to social media research.
population of Twitter users might systematically over-represent PTSD incidence, this sample could be justified as consisting mostly of individuals without PTSD, based on simple probability. If any Twitter users included in the comparison sample actually did have PTSD, it could reasonably be assumed that their “PTSD signal” would be drowned out by the overwhelming healthy signal, due to the relative magnitude of the comparison sample. Similar concerns about bulk ingestion data contamination apply even for those individuals who made explicit disclosures of PTSD diagnosis. Without any reliable way to date the traumatic event that caused PTSD, it is possible that some of the tweets included in the PTSD sample were posted by individuals prior to their PTSD-causing trauma, effectively meaning that healthy tweets are included in the PTSD sample.

While cross-contamination is problematic from a quality control perspective, it becomes a major concern only in the case that an analysis yields no differences between affected and healthy samples. Cross-contamination should only serve to dilute a true difference, suggesting that analytical results showing an observable difference between samples would only be stronger if contamination issues were resolved. As such, this concern is less salient when results are strong.

A third category of limitations to the bulk ingestion approach concerns sample bias. The set of individuals who have been diagnosed with depression and choose to actively and directly disclose this information on Twitter is almost certainly a subset of the total set of individuals who use Twitter and have been diagnosed with depression. Individuals who choose to disclose personal health information in public may be markedly different from individuals who keep such information private, both in psychological profile, and also perhaps in the way they present depressive symptoms. As such, any inferences drawn from bulk ingestion analyses cannot
reliably extend even to other depressed Twitter users (not to mention extension to the larger depressed population beyond Twitter - an issue which I return to later). An additional problem with this disclosure bias is that we can be sure that the entire set of non-disclosing depressed Twitter users is potentially included in the user pool from which the comparison sample is drawn, lending further emphasis to the cross-contamination concerns detailed above.

The lack of specificity regarding diagnosis (and, in the case of PTSD, trauma dates) made bulk ingestion suboptimal for my research project, as I was specifically interested in social media activity relative to these dates. Given enough data, it might have been possible to identify enough Twitter accounts which not only tweeted explicit diagnosis disclosures, but also explicit dates of diagnosis, to allow for a sufficient sample size for analysis. In my initial research, however, I searched through an archive of several hundred million tweets, and found very few cases which clearly disclosed a specific diagnosis date, although several hundred did provide explicit disclosure of diagnosis. I initially thought that a solution to the bulk ingestion non-specificity would be to send study participation invitations to the account holders who disclosed diagnosis, but not date of diagnosis, but Twitter’s anti-spam filters prevented me from pursuing that recruitment avenue.

Social media analyses based on bulk ingestion methods are also limited in that they rely on scanning text to identify relevant content. In those cases where social media platforms consist primarily of visual content, such as Instagram, bulk ingestion methods fail in the absence of a more appropriate way to search through the mass of data collected. Additionally, many social media platforms, including Instagram, do not provide a streaming public API, but rather limit data collection to individual user accounts, for which the requester must provide a specific account username. As such, bulk ingestion is not a feasible data collection strategy for these
platforms, although given enough time, and provided a researcher has access to a sufficient list of usernames, it is technically possible to amass a sizeable data set via publicly available means. All of these issues prompted me to pursue a strategy of active recruitment, in which data collection involved engaging with study participants directly.

*Active Recruitment*

Active recruitment is a more conventional data collection approach in behavioral research, in which study participants are personally asked to provide information about themselves. The immediate problem facing an active recruitment approach to data collection in this context is one of participant availability. Consider the extremely narrow defining characteristics of the target population of interest (in the case of depressed individuals):

- Has been diagnosed with depression (6.7% of US adult population) (Kessler et al., 2005)
- Actively used Instagram/ Twitter at and around time of diagnosis (21% of US population uses Twitter in 2016) (Smith, 2016)
- Willing to divulge depression history to academic researchers
- Willing to share Instagram/ Twitter account history to academic researchers

It is reasonable (and quite correct, as it turns out) to assume that most traditional recruitment avenues would be unhelpful for finding participants that fit this profile. Informal polling of Harvard undergraduates suggested that very few, if any, students would be willing to share their personal social media accounts with researchers. Efforts to contact depression support groups in major cities across the United States did not yield any willing participants. In the latter case of contacting support groups, while some groups had specific policies against

---

5 In subsequent chapters I address the privacy and ethical concerns which arise with any kind of analysis linking social media profiles to personal health information; for now, suffice to say that it would have been an inefficient use of available resources to pursue on-campus recruitment.
working with research teams, others were unable to be contacted for reasons similar to Twitter’s anti-spam filters rejecting my invitations - contact information was available online only, and often accessible only through the medium of community-organizing websites such as MeetUp.com. In these instances, invitations were blocked by non-solicitation filters. I created a website, SocialHealthStudy.com, where interested parties could learn about my research, determine whether they were qualified to participate, and which linked directly to consent and data collection materials. Efforts to increase site visibility and visitor traffic never really took off, however, and ended up as another recruitment dead end.

Amazon’s Mechanical Turk (MTurk) crowdwork platform provided a suitable avenue for my recruitment efforts. All participants in all studies were recruited from MTurk. I used Turk Prime, an Amazon affiliate, to manage recruitment logistics. In particular, Turk Prime solved a privacy concern associated with conducting my research on MTurk. MTurk terms of use require that Requesters (people, like me, who request that Workers complete a task) post no tasks which run the risk of exposing Workers’ true identities - in other words, tasks are prohibited which link an MTurk Worker ID with that Worker’s personal identity. While the surveys I created did not ask for personal identifiers, the fact that they also involved sharing social media data meant that Worker IDs could plausibly be associated with Worker identities. Fortunately, Turk Prime offers a service to Requesters which anonymizes Worker IDs, so that any data collected is only associated with an anonymous, randomized character string. Turk Prime allowed me to use MTurk without violating their terms of use; future research in this vein which uses MTurk for recruitment should take care to employ a similar service.

Using MTurk for recruitment further narrowed the population size which I wanted to sample. In addition to the four criteria listed above (received diagnosis, used social media,
willing to share health data, willing to share social media), participants also had to be actively working on MTurk, shrinking both the available pool of potential participants, as well as the ecological validity of my findings\(^6\). On the other hand, MTurk allows Requesters to restrict access to task advertisements based on Worker quality ratings, which allows for some measure of data quality control that can be more difficult to achieve in in-person settings.

Reluctance to share social media data was, by far, the primary limiting factor in data collection. Several hundred MTurk workers (almost a thousand, in the case of the Twitter/depression study) began work on the surveys for each study, read and accepted the consent materials, and answered questions relating to inclusion criteria. Almost all were willing to share personal health history, but over 40% of participants refused to share their social media data, even after being given assurances that it would be handled by a team of not-for-profit, academic researchers who were dedicated to keeping their data safe and private. This high dropout rate associated with refusal to share social media data may be related to the string of recent revelations that social media research - conducted both by companies and academic teams - had either manipulated or exposed user data without consent (Hackett, 2016; Kirkegaard & Bjerrekaer, 2016; Kramer, Guillory & Hancock, 2014; Rudder, 2014). Establishing trust with potential research participants that their data will be handled ethically and in a secure manner is a matter of ongoing concern for social computational scientists\(^7\).

Most of the information collected from participants came from surveys powered by the Qualtrics survey platform, which is a common survey software that bears no special explication here. This research project, however, required connecting Qualtrics both to an external data

---

\(^6\) Validity issues are discussed in more detail in subsequent chapters.

\(^7\) Consider the many extant studies which have used bulk ingestion to collect social media data; it is unclear that, even given the technically “public” nature of these data, individual social media users would feel comfortable with their information being used in this way.
collection method (the apps that collected social media data from either Twitter or Instagram) and to a database on my own server (for verification and storage). These proved to be unusual and relatively difficult tasks, and I have allocated a section in the Appendix (see Appendix: Hacking Qualtrics) to briefly describe the technical processes involved in the case that readers may be interested in creating a similar setup.

**Step 3: Quality control**

Several filtering procedures were applied, in addition to the basic inclusion criteria presented to all participants, in order to ensure data quality as best as possible. MTurk data can be unreliable, as workers can earn money for completing tasks, regardless of the quality of data they provide. To address this, I restricted all MTurk task advertisements so that they were only viewable to workers who had a 95% approval rating or higher, and who had completed at least 100 previous MTurk tasks. MTurk workers with these quality ratings have been found to provide reliable, valid survey responses (Peer, Vosgerau, & Acquisiti, 2013). Task access was also restricted to U.S. IP addresses, as MTurk data collected from outside the United States are generally of poorer quality (Litman, Robinson, & Rosenzweig, 2014). All workers were only permitted to take the survey once.

In addition to placing MTurk restrictions, I also excluded some participant data, post-collection. Accounts with fewer than five posts across an entire user’s history were excluded from analysis\(^8\). I also excluded participants who scored below clinical thresholds on the scales used in the depression and PTSD studies. This exclusion decision was in keeping with the precedent set by De Choudhury et al. (2013). Together, these quality control measures helped remove some of the noise that may have masked predictive signals of interest in collected data.

---

\(^8\) By way of context, the median number of posts for Instagram users in the collected sample was 122.5, and between 862 to 946 for Twitter, depending on the condition being studied.
Step 4: Feature extraction

The tasks of choosing salient predictive features in data, and finding effective methods of extracting those features from data, are arguably the most vital challenges in any machine learning project - much more so than the choice of learning algorithm. This is different from most behavioral experiments, in which independent variables, based on purposeful manipulations, along with any desired covariates, form the feature set. In this research project, all studies were observational in nature. The design challenge, therefore, lay not in designing a good manipulation to test a hypothesis, but in finding the right features to pull out of the raw data, which could reveal psychological signals of interest and accurately predict target outcomes. The psychological component of this step in the research design lies in the motivation behind each choice of predictive feature; the computational component is found in the means of feature extraction.

Both aspects of the feature extraction process are explained at length in the study-specific chapters that follow, although given the space and topical constraints of this report I chose to omit an extensive explication of the programming specifics involved. I would be remiss, however, in failing to note the sizable effort that went into architecting and developing the code base behind this project, especially for the sake of psychologists and other social scientists interested in similar lines of research. Many of the predictive features used in my research required their own smaller projects to develop. For example, consider the count of human faces in photographs, described in Chapter III. With over 40,000 photographs to evaluate, it would have been impractical to employ humans to count the faces in each photograph. Instead, some kind of automated facial detection program was necessary to generate this feature. I was unable to find any commercial face detection software which afforded the necessary privacy guarantees,
given the sensitive human subjects data involved in my work, so I developed my own face detection algorithm based on a number of open-source code samples found online. The algorithm, in turn, needed to be vetted for internal validity, and required several reformulations before it achieved a satisfactory level of accuracy in detecting human faces in photographs. Despite the current social buzz surrounding artificial intelligence, machine learning, and other advanced technologies, there still are relatively few tools available to researchers which are ready “out of the box” for human subjects research.

The technical details in the rest of this report will remain, for the most part, limited to methodological considerations concerning study design and analytic strategies. This decision is motivated, in part, by my interest in focusing on the value of this report from a social and psychological scientific perspective, but it also reflects the sheer volume of the programmatic backbone of this project. In order to collect, transform, store, and analyze the data collected for my project, I wrote roughly 6,000 lines of code (considerably more than that when false starts and dead end methods are accounted for), spanning five programming languages and protocols. To put that quantity in context, if those lines were added to my report, they would increase the length of this document by 286 double-spaced pages. Instead, I have posted the entirety of the code base I developed, along with extensive documentation, as an open-source project (see github.com/andrewreece/predicthealth).

**Step 5: Analysis and modeling**

The primary analytical goal of this project was to train a statistical model, capable of differentiating social media content generated by individuals with and without a specific health condition, such as depression. Specific details for each study are presented in the full-report
chapters that follow. Here, I address the key high-level choices I made which influenced all analyses, along with some justification for their relative merits.

Binary classification

Models were trained using a binary response variable, which labeled each observation as having or not having the health condition of interest (depression, PTSD, etc.). This restricted the set of possible modeling choices to classification algorithms, such as logistic regression, support vector machines, and decision tree methods. Models based on continuous responses would have been feasible in those cases where the health condition of interest was measurable with a scale (e.g. depression, PTSD), but a continuous-response approach would have excluded conditions which are not marked by continuous scales. As the goal was to establish a robust methodology, capable of applying to any number of conditions, a binary classification schema was most appropriate.

Accuracy metrics

When evaluating the results of a binary classifier, it is important to choose an appropriate loss function. This is especially true when data exhibit a class imbalance, as was often the case in the studies reported here (i.e., it was often easier to find more participants without a given health condition than with it). For example, interpreting classification accuracy based on a simple proportion of correct classifications to total classifications (called raw or naive accuracy) is not only misleading in the case of imbalanced classes, but is also opaque to the specific strengths and weaknesses of a binary classifier. Binary classification generates four categories of results: “true negative” (model predicts healthy class correctly), “true positive” (model predicts affected class correctly), “false negative” (model predicts healthy class incorrectly), and “false
positive” (model predicts affected class incorrectly). A confusion matrix is typically used to display these classification quadrants (see Figure 1).
<table>
<thead>
<tr>
<th></th>
<th>Predicted Healthy</th>
<th>Predicted Affected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Known Healthy</td>
<td>True Negative</td>
<td>False Positive</td>
</tr>
<tr>
<td>Known Affected</td>
<td>False Negative</td>
<td>True Positive</td>
</tr>
</tbody>
</table>

Figure 1. Labeled cells of a confusion matrix.

More informative approaches to classification accuracy incorporate information from each of these four categories. In evaluating classifier accuracy, I report five common metrics: precision (ratio of true positives to predicted positives), recall (ratio of true positives to actual positives), specificity (ratio of true negatives to total negatives), negative predictive value (ratio of true negatives to predicted negatives), and \( F_1 \) score. An \( F_1 \) score ranges from 0 to 1, with higher scores indicating better accuracy, and is computed as the harmonic mean of precision and recall:

\[
F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}
\]

An additional means of considering binary classification accuracy is via a Receiver Operating Characteristic (ROC) curve\(^9\). ROC curves plot a prediction vector’s true positive rate against its false positive rate. See Figure 2 for an example ROC curve plot. The diagonal dashed line represents the case where there are just as many false positives as true positives (in other words, no predictive power is observable above chance). The larger the area under the ROC curve, the better the ratio of true positives to false positives. A perfect prediction vector would hug the left and top edges of the graph. Both ROC curves and the suite of accuracy metrics described above are used to evaluate classifier accuracy.

\(^9\) Davis & Goadrich (2006) show that an ROC curve is closely related to precision and recall statistics, and serves as a close proxy except in cases of severe class imbalance.
Testing practical significance with holdout data

Statistical significance does not always imply practical significance. In common frequentist inferential statistics employed in social science research, various measures of effect size are used to estimate the practical significance of independent variables (see Rosenthal, 1994). Among machine learning and computational scientists, it is more common to measure inferential power directly via a train/test paradigm. In this design, a portion of the data to be analyzed (often in the range of 20-50%) is omitted when fitting a statistical model. The data used to fit a model is called the training set and the omitted data is called the test set. Once a model has been fit with the training data, the observations of the test set (excluding the response variable values) are used to generate estimates on the response variable. The true response values in the test set are compared with model estimates using the appropriate loss function (in
the case of classification: precision, recall, F1, etc.) to evaluate the quality of the model. This train/test paradigm allows for the immediate and direct assessment of model quality, instead of using a test statistic (e.g. Cohen’s $d$) as a proxy. The train/test procedure is used less frequently in social science research, often because sample sizes are small to begin with, and further reducing the available data with which to train a model may be infeasible. Computational science research usually involves larger datasets, and so can afford this reduction in training sample size. All the analyses reported here employ a train/test paradigm to directly assess the statistical power of predictive models.

Algorithm and modeling selection

Specific algorithmic choices are described in more depth in the Instagram and Twitter study chapters, respectively. Here, I present some justification for the strategic use of different classes of learning algorithms for each of the social media platforms studied.

Computational analysis of photographic social media data, with the aim of extracting meaningful psychological insights about users’ mental health, was an entirely novel proposal at the time I planned this project. In debuting a new methodology, it is especially useful to be able to provide some interpretative context to be able to answer a first natural question: Why should this new approach be carried out, and what does it mean if it works? In the domain of machine learning, however, the most powerful results are not always the most interpretable results. If pure predictive power were the only goal, then it might not matter as much how the inner workings of an algorithmic decision process play out. A psychological scientist, however, might care more about the driving influences behind a strong model. As this work straddles the fields

---

10 In those cases where variable selection is performed after an initial model fit, it is common to create three partitions of the data: a training set, a test set, and a validation set. The test set may be used in exploratory data analysis and in the model selection process, but the validation set remains completely unexamined until a final model has been determined and deemed ready for testing. For the studies presented in this report, no variable selection took place and so a simple train/test setup was sufficient.
of computational and psychological science, I attempted to find a balance between maximizing predictive accuracy, and focusing on the explanation behind the accuracy.

Consider the Random Forests classification algorithm, which ended up showing the strongest performance in all studies presented here. In the Random Forests implementation I used (from Python’s scikit-learn module), a measure of feature importance is available, but the directionality of each predictor’s impact on the classification process is not (Pedregosa et al., 2011). For example, I may learn from the model output that the brightness of a photograph is an important feature in predicting depression, but Random Forests will not tell me whether it is an increase, or a decrease, in brightness that is associated with an increased likelihood of depression.

As an accompaniment to the relatively opaque Random Forests output, I also fit a Bayesian logistic regression to the Instagram data. Logistic regression, while it was outperformed by Random Forests in terms of predictive accuracy, still largely agreed with the Random Forests model’s assessment of feature importance. The added benefit of regression output is that it provides directional indicators, along with magnitude of impact, per feature. I presented both logistic regression and Random Forests output together in the Instagram study, along with a number of other contextual analyses, to frame the methodology in a narrative that was both intuitive and helpful to other researchers.

In addition, I asked human raters to evaluate the photos collected, on a number of common affective categories. While collecting ratings data contradicted a main goal of this research - to provide a fast, efficient, and computational approach for predictive health screening - it did provide useful context for the performance of the primary computational model. Building a separate predictive model, based only on human ratings, allowed only for a comparison of
model strength between computational and behavioral features, but it also allowed for a
correlational analysis between feature sets. Determining whether the selected computational
features aligned with common human sentiments helped shed light on another natural question
arising from this methodology, namely, “Are computational features and human perceptions
picking up on the same signals?” This point is discussed in greater depth in the Instagram
chapter that follows.

In the case of Twitter data, at the time I began my research a small number of studies had
already been published showing that psychological markers could be detected via computational
analysis (Coppersmith et al, 2014; De Choudhury et al, 2013; De Choudhury et al., 2014). In
particular, De Choudhury et al. (2013) established a methodological approach, incorporating
active recruitment and predictive modeling, which closely reflected my own research goals. As
such, the Twitter arm of my research was intended as a different sort of contribution from the
novelty of Instagram. First, it served as an independent replication of De Choudhury et al.
(2013). Second, it improved on a number of methodological details from De Choudhury et al.
(2013), including, most notably, an analysis of whether signals of mental illness were detectable
before affected individuals had received a first diagnosis of a given condition. Third, I
developed a novel statistical analysis using state-space modeling, that tracked the probabilistic
rise and fall of target health conditions in time - something no social media research had
attempted to do previously. For these reasons, the choice of statistical techniques in the Twitter
studies reported here are focused somewhat more on improving and innovating new methods,
rather than on easily accessible interpretation. (The graph output of the state-space time series
modeling is a notable exception to this trend.)
Units of observation\(^{11}\)

Determining the best time span for analysis raises a difficult question: When and for how long does mental illness occur? Receiving a clinical diagnosis of depression or PTSD does not imply that an individual remains in a persistent state of illness. Even in the case of pregnancy, where the classification of “pregnant” does not change from conception to giving birth, the intensity of experience associated with pregnancy can change dramatically. In this light, to conduct analysis with an individual’s entire posting history as a single unit of observation is a dubious proposition in general, and plainly incorrect in those cases where individuals made posts prior to the time of their developing a given health condition. At the other extreme, to take one tweet or photograph as a unit of observation runs the risk of being too granular.

DeChoudhury et al. (2013) looked at all of a given user’s tweets in a single day, and aggregated those data into per-person, per-day units of observation. In this report we have followed the convention of aggregated “user-days” as a primary unit of analysis, rather than try to categorize a person’s entire history, or analyze each individual post\(^{12}\). In the case of Twitter data, it has been observed that many Twitter users do not generate enough daily content to make for robust unigram sentiment analysis (Bliss et al., 2012). For completeness, we conducted analyses using both daily and weekly units of observation\(^{13}\). In the case of Instagram data, which does not use any kind of text analysis, concerns regarding unigram analysis are obviated, and user-day analyses were conducted.

---

\(^{11}\) This discussion is revisited in both of the study-specific chapters to follow.

\(^{12}\) Occasionally, when reporting results we refer to observations or tweets as “depressed”, e.g. “depressed tweets received fewer likes”. It would be more correct to use the phrase “tweet data from depressed participants, aggregated by user-days” instead of “depressed tweets”, but we chose to sacrifice a degree of technical correctness for the sake of clarity.

\(^{13}\) We also considered hourly units of observation, as De Choudhury et al. (2013) found substantial differences in diurnal patterns between depressed and healthy subjects. In our sample, however, exploratory analysis showed no differences in hourly posting trends between these groups, and we did not conduct analyses using hourly units.
Study presentation

Data collection for the Instagram studies progressed at a much slower pace than for the Twitter studies. Only the depression condition collected data from enough viable participants to provide stable analytical results. Further rounds of Instagram data collection were blocked by a change in the Instagram API terms of use, which invalidated all existing applications in June 2016. Given time constraints on this project, in addition to the fact that the depression data were sufficient to conduct a full analysis, I limited the Instagram arm to the depression condition only.

In the case of Twitter, sufficient sample sizes were collected for both the depression and PTSD conditions. As the analytical methodology is the same for all conditions studied using Twitter data, and given the relatively high comorbidity rates of depression and PTSD, I reported findings from these two conditions in a single chapter. The two chapters that follow are the reports of findings from the Instagram/depression, and Twitter/depression/PTSD studies, respectively.
Note: I collaborated with other researchers on the projects described in Chapters III and IV. Accordingly, I switch to use of the singular plural, “we”, for these two chapters, to reflect this fact. The writing, as well as the large majority of design and analysis, is my own, although modified versions of each chapter, not presented here, were edited and augmented by my co-authors prior to submission for publication. C. Danforth is my co-author for the project described in Chapter III, and A. Reagan, K. Lix, P. Dodds, C. Danforth, and E. Langer are my co-authors for the project described in Chapter IV. Any errors or omissions in the versions of the reports I have presented here are entirely my own.
Chapter III. Depression and Instagram

The advent of social media presents a promising new opportunity for early detection and intervention in psychiatric disorders. Predictive screening methods have successfully analyzed online media to detect a number of harmful health conditions (Christakis & Fowler, 2010; Coppersmith et al., 2014; De Choudhury et al., 2013a; De Choudhury et al., 2014; De Choudhury et al., 2013b; De Choudhury et al., 2016; Katikalapudi et al., 2012; Moreno et al., 2011; Moreno et al., 2012; Paparrizos et al., 2016; Schmidt, 2012). All of these studies relied on text analysis, however, and none have yet harnessed the wealth of psychological data encoded in visual social media, such as photographs posted to Instagram. In this report, we introduce a methodology for analyzing photographic data from Instagram to predictively screen for depression.

There is good reason to prioritize research into Instagram analysis for health screening. Instagram members currently contribute almost 100 million new posts per day (Instagram, 2016), and Instagram’s rate of new users joining has recently outpaced Twitter, YouTube, LinkedIn, and even Facebook (Chaffey, 2016). A nascent literature on depression and Instagram use has so far either yielded results that are too general or too labor-intensive to be of practical significance for predictive analytics (Andalibi, Ozturk & Forte, 2015; Lup, Trub, & Rosenthal, 2015). In our research, we incorporated an ensemble of computational methods from machine learning, image processing, and other data-scientific disciplines to extract useful psychological indicators from photographic data. Our goal was to successfully identify and predict markers of depression in Instagram users’ posted photographs.

Hypothesis 1: Instagram posts made by individuals diagnosed with depression can be reliably distinguished from posts made by healthy
controls, using only measures extracted computationally from posted photos and associated metadata.

**Photographic markers of depression**

Photographs posted to Instagram offer a vast array of features that might be analyzed for psychological insight. The content of photographs can be coded for any number of characteristics: Are there people present? Is the setting in nature or indoors? Is it night or day? Image statistical properties can also be evaluated at a per-pixel level, including values for average color and brightness. Instagram metadata offers additional information: Did the photo receive any comments? How many “Likes” did it get? Finally, platform activity measures, such as usage and posting frequency, may also yield clues as to an Instagram user’s mental state. We incorporated only a narrow subset of possible features into our predictive models, motivated in part by prior research into the relationship between mood and visual preferences.

In studies associating mood, color, and mental health, healthy individuals identified darker, grayer colors with negative mood, and generally preferred brighter, more vivid colors (Barrick, Taylor, & Correa, 2002; Boyatzis & Varghese, 1994; Carruthers, Morris, Tarrier, & Whorwell, 2010; Hemphill, 1996). By contrast, depressed individuals were found to prefer darker, grayer colors (Carruthers et al., 2010). In addition, Barrick, Taylor, & Correa (2002) found a positive correlation between self-identification with depression and a tendency to perceive one’s surroundings as gray or lacking in color. These findings motivated us to include measures of hue, saturation, and brightness in our analysis. We also tracked the use of Instagram filters, which allow users to modify the color and tint of a photograph.

Depression is strongly associated with reduced social activity (American Psychiatric Association, 2000; Bruce & Hoff, 1994). As Instagram is used to share personal experiences, it is
reasonable to infer that posted photos with people in them may capture aspects of a user’s social life. On this premise, we used a face detection algorithm to analyze Instagram posts for the presence and number of human faces in each photograph. We also counted the number of comments and likes each post received as measures of community engagement, and used posting frequency as a metric for user engagement.

**Early screening applications**

Hypothesis 1 is a necessary first step, as it addresses an unanswered basic question: Is depression detectable in Instagram posts? On finding support for Hypothesis 1, a natural question arises: Is depression detectable in Instagram posts, *before the date of first diagnosis*? After receiving a depression diagnosis, individuals may come to identify with their diagnosis (Cornford, Hill, & Reilly, 2007; Karp, 1994). Individuals’ self-portrayal on social media may then be influenced by this identification. It is possible that a successful predictive model, trained on the entirety of depressed Instagram users’ posting histories, might not actually detect depressive signals, per se, but rather purposeful content choices intended to convey a depressive condition. Training a model using only posts made by depressed participants prior to the date of first diagnosis addresses this potential confounding factor.

Hypothesis 2: Instagram posts made by depressed individuals prior to the date of first clinical diagnosis can be reliably distinguished from posts made by healthy controls.

If support is found for Hypothesis 2, this would not only demonstrate a methodological advance for researchers, but also serve as a proof-of-concept for future healthcare applications. As such, we benchmarked the accuracy of our model against the ability of general practitioners to correctly diagnose depression as shown in a meta-analysis by Mitchell, Vaze, and Rao (2009).
The authors analyzed 118 studies that evaluated general practitioners’ abilities to correctly diagnose depression in their patients, without assistance from scales, questionnaires, or other measurement instruments. Out of 50,371 patient outcomes included across the pooled studies, 21.9% were actually depressed, as evaluated separately by psychiatrists or validated interview-based measures conducted by researchers. General practitioners were able to correctly rule out depression in non-depressed patients 81% of the time, but only diagnosed depressed patients correctly 42% of the time. We refer to these meta-analysis findings as a comparison point to evaluate the usefulness of our models.

A major strength of our proposed models is that their features are generated using entirely computational means - pixel analysis, face detection, and metadata parsing - which can be done at scale, without additional human input. It seems natural to wonder whether these machine-extracted features pick up on similar signals that humans might use to identify mood and psychological condition, or whether they attend to wholly different information. A computer may be able to analyze the average saturation value of a million pixels, but can it pick out a happy selfie from a sad one? Understanding whether machine learning and human opinion are sensitive to the same indicators of depression may be valuable information for future research and applications. Furthermore, insight into these issues may help to frame our results in the larger discussion around human versus machine learning, which occupies a central role in the contemporary academic landscape.

To address these questions, we solicited human assessments of the Instagram photographs we collected. We asked new participants to evaluate photos on four simple metrics: happiness, sadness, interestingness, and likability. These ratings categories were intended to capture human impressions that were both intuitive and quantifiable, and which had some
relationship to established depression indicators. DSM-IV criteria for Major Depressive Disorder includes feeling sad as a primary criterion, so sadness (and its anti-correlate, happiness) seemed obvious candidates as ratings categories (American Psychiatric Association, 2000). Epstein et al. (2010) found depressed individuals “had difficulty reconciling a self-image as an ‘outgoing likeable person’”, which prompted likability as an informative metric. We hypothesized that human raters should find photographs posted by depressed individuals to be sadder, less happy, and less likable, on average. Finally, we considered interestingness as a novel factor, without a clear directional hypothesis.

Hypothesis 3a: Human ratings of Instagram posts on common semantic categories can distinguish between posts made by depressed and healthy individuals.

Hypothesis 3b: Human ratings are positively correlated with computationally-extracted features.

If human and machine\textsuperscript{14} predictors show positive correlation, we can infer that each set of features tracks similar signals of depression. In this case, the strength of the human model simply suggests whether it is better or worse than the machine model. On the other hand, if machine and human features show little or no correlation, then regardless of human model performance, we would know that the machine features are capable of screening for depression, but use different information signals than what are captured by the affective ratings categories.

**Method**

\textsuperscript{14} The term “machine” (e.g. “machine predictors”, “machine model”) is used as shorthand for the computational feature extraction process we employed. Significant human biases informed this process, however, as the initial selection of features for extraction involved entirely human decision-making.
This study was reviewed and approved by the Harvard University Institutional Review Board, approval #15-2529 and by the University of Vermont Institutional Review Board, approval #CHRMS-16-135. All study participants were informed of and acknowledged all study goals, expectations, and procedures, including data privacy procedures, prior to any data collection. Surveys were built using the Qualtrics survey platform (www.qualtrics.com). Analyses were conducted using the Python (v2.7) and R (v3.3) programming languages (Python Software Foundation, 2016; R Core Team, 2016). Social media data collection apps were written in Python, using the Instagram developer’s Application Programming Interface (API) (www.instagram.com/developer).

**Data Collection**

Data collection was crowdsourced using Amazon’s Mechanical Turk (MTurk) crowdwork platform. In an effort to reduce noisy and/or unreliable data, we employed several MTurk quality assurance measures. Our surveys were only visible to MTurk crowdworkers who had completed at least 100 previous tasks with a minimum 95% approval rating; MTurk workers with this level of experience and approval rating have been found to provide reliable, valid survey responses (Peer, Vosgerau, & Acquisti, 2013). We also restricted access to only American IP addresses, as MTurk data collected from outside the United States are generally of poorer quality (Litman, Robinson, & Rosenzweig, 2014). All participants were only permitted to take the survey once. We also excluded participants who had successfully completed our survey, but who had a lifetime total of fewer than five Instagram posts.

---

15 The number five here was chosen principally as a filter against a few observed cases where users actually created Instagram accounts, seemingly for the express purpose of qualifying for our study (and to collect the associated compensation). In these few cases, participants had only zero, one, or two total posts, and so a cutoff of five was sufficient to exclude these cases.
Separate surveys were created for depressed and healthy individuals. In the depressed survey, participants were invited to complete a survey that involved passing a series of inclusion criteria, responding to a standardized clinical depression survey, answering questions related to demographics and history of depression, and sharing social media history. The inclusion criteria ensured that participants: (1) Had received a first formal diagnosis of depression sometime between 2013 to present\textsuperscript{16}, (2) Had been active users of Instagram at and around the time of diagnosis, (3) Could recall the exact date of their first depression diagnosis, and (4) Were willing to share Instagram history with our research team for analysis. Participants who did not meet all of these criteria were barred from further participation and excluded from analysis.

We used the CES-D (Center for Epidemiologic Studies Depression Scale) questionnaire to screen participant depression levels (Radloff, 1977). CES-D assessment quality has been demonstrated as on-par with other depression inventories, including the Beck Depression Inventory and the Kellner Symptom Questionnaire (Fountoulakis, 2007; Zich, Attkisson, & Greenfield, 1990). Participants with CES-D scores of 21 or lower were excluded from analysis, following the precedent of De Choudhury et al. (2013). Previous studies have indicated that a CES-D score of 22 represents an optimal cutoff for identifying clinically relevant depression across a range of age groups and circumstances (Cuijpers, Boluijt, & van Straten, 2007; Haringsma et al., 2004).

Healthy participants were screened to ensure no history of depression and active Instagram use. The survey for healthy participants collected age and gender data from participants. All qualified participants were asked to share their Instagram usernames and history. An app embedded in the survey allowed participants to securely log into their Instagram\textsuperscript{16}

\textsuperscript{16} The 2013 boundary was determined by informally estimating the time period that a) participants could reasonably remember exact diagnosis dates, and b) Instagram user growth showed a dramatic increase, raising the likelihood of our finding Instagram users as participants (Statista, 2016).
accounts and agree to share their data\textsuperscript{17}. Upon securing consent, we made a one-time collection of participants’ entire Instagram posting histories. In total we collected 43,950 photographs from 166 Instagram users.

We asked a different set of MTurk workers to rate the Instagram photographs collected. This new task asked participants to rate a random selection of 20 photos from the data we collected. Raters were asked to judge how interesting, likable, happy, and sad each photo seemed, on a continuous 0-5 scale. Each photo was rated by at least three different raters, and ratings were averaged across raters. Raters were not informed that photos were from Instagram, nor were they given any information about the study participants who provided the photos, including mental health status. Each ratings category showed good inter-rater agreement (see Results for more on agreement measurement).

Only a subset of participant Instagram photos were rated (N=13,184). We limited ratings data to a subset because this task was time-consuming for workers, and so proved a costly form of data collection. For the depressed sample, ratings were only made for photos posted within a year in either direction of the date of first depression diagnosis. Within this subset, for each user the nearest 100 posts prior to the diagnosis date were rated. For the control population, the most recent 100 photos from each user’s date of participation in this study were rated.

**Participant safety and privacy**

Data privacy was a concern for this study. Strict anonymity was nearly impossible to guarantee to participants, given that usernames and personal photographs posted to Instagram often contain identifiable features. We made sure participants were informed of the risks of being personally identified, and assured them that no data with personal identifiers, including usernames, would be made public or published in any format.

\textsuperscript{17} Data collection source code is available on Github, github.com/andrewreece/predicthealth
This study design raised two important issues regarding ethical research practices, as it concerned both individuals with mental illness and collection of potential personal identifiers. We took a number of steps to address these concerns. Regarding participants’ mental health, we informed participants that this study did not offer any treatment for depression, and was not intended to replace or improve upon any existing treatment. We did, however, provide the phone number of a national counseling hotline, as well as a link to the Depression and Bipolar Support Alliance, and we encouraged participants to engage these resources if they felt in need of assistance.

Regarding data privacy, we enacted several measures to treat participants’ respectfully and ethically. While this is always an expected standard in human subjects research, we felt it especially important to emphasize to participants, given that recent social media studies have caused notable concern due to participants’ data either having been exposed or manipulated without explicit consent (Hackett, 2016; Kirkegaard & Bjerrekaer, 2016; Kramer, Guillory & Hancock, 2014; Rudder, 2014). In our own research, strict anonymity was nearly impossible to guarantee to participants, given the nature of the data collected. Usernames and personal photographs posted to Instagram are often inherently specific to participants’ identities (e.g. photo “selfies”, or usernames containing actual names). As we potentially had the capacity to personally identify study participants (and further, to link their identities to sensitive health information), we endeavored to ensure that participants were well-informed about the potential privacy risks involved in participation.

Study participants were informed of the risks of being personally identified from their social media data, and assured them that no personal identifiers, including usernames, would ever be made public or published in any format. We used Turk Prime, an interface for
conducting MTurk studies, to mask participants’ MTurk worker IDs from us as an extra added layer of privacy, and we informed participants of this. All participants were informed both by the MTurk HIT advertisement and during the survey itself that any photos posted to Instagram which they chose to share might be viewed and rated by other people, as well. We made it clear that any links between social media data and private personal health data would be available only to our team of researchers, and we informed participants that they may request to have their data removed at any time.

**Feature extraction**

Several different types of information were extracted from the collected Instagram data. We used total posts per user, per day, as a measure of user activity. We gauged community reaction by counting the number of comments and “likes” each posted photograph received. Pixel-level averages were computed for Hue, Saturation, and Value (HSV), three color properties commonly used in image analysis. Hue describes an image’s coloring on the light spectrum (ranging from red to blue/purple). Lower hue values indicate more red, and higher hue values indicate more blue. Saturation refers to the vividness of an image. Low saturation makes an image appear grey and faded. Value refers to image brightness. Lower brightness scores indicate a darker image. See Figure 3 for a comparison of high and low HSV values. We also checked metadata to assess whether an Instagram-provided filter was applied to alter the appearance of a photograph. Collectively, these measures served as the feature set in our primary model. For the separate model fit on ratings data, we used only the four ratings categories (happy, sad, likable, interesting) as predictors.
Figure. 3. Comparison of HSV values. Right photograph has higher Hue (bluer) and lower Brightness (darker) than left photograph. Instagram photos posted by depressed individuals had HSV values shifted towards those in the right photograph, compared with photos posted by healthy individuals.

Face detection software was used to determine whether or not a photograph contained a human face, as well as count the total number of faces in each photo, as a proxy measure for participants’ social activity levels. We adapted an elementary face detection script, based on open source code (Nguyen, 2015).  

Face detection algorithm accuracy was assessed by manually coding a random sample of 400 photos (100 photos from each of combination of depressed/healthy, detected/undetected). Detection accuracy was roughly equal across groups (Table 1).

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Face detection algorithm accuracy across groups and detection categories</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No face</td>
</tr>
<tr>
<td>Healthy</td>
<td>79%</td>
</tr>
<tr>
<td>Depressed</td>
<td>77%</td>
</tr>
</tbody>
</table>

Note. Algorithm performance was roughly equal for healthy and depressed groups in each detection category.

18 The main adjustment from the open source code was to run the detection loop twice, using two differing scale factors. A single scale factor had difficulty finding both small and large faces. Parameters used: scale_factors = 1.05, 1.4, min_neighbors = 4, min_size = (20px,20px)
The mean difference in counted faces (detected faces minus actual faces), indicated that the algorithm slightly undercounted the number of faces in photos, for both depressed participants ($\mu = -0.015, \sigma = 1.21$) as well as healthy participants ($\mu = -0.215, \sigma = 2.07$). In both groups, the algorithm undercounted by less than a single face, on average. Despite the algorithm mediocre accuracy rates, we felt comfortable in applying its results for our purposes, given that our main concerns were 1) that the algorithm was reasonably accurate, i.e. correct predictions above chance, and 2) that the algorithm was about as accurate for both groups (depressed and healthy). Both of these conditions were satisfied; future iterations of this research would do well to build a more accurate face detection program.

Units of observation

In determining the best time span for this analysis, we encountered a difficult question: When and for how long does depression occur? A diagnosis of depression does not indicate the persistence of a depressive state for every moment of every day, and to conduct analysis using an individual’s entire posting history as a single unit of observation is therefore rather specious. At the other extreme, to take each individual photograph as units of observation runs the risk of being too granular. DeChoudhury et al. (2013a) looked at all of a given user’s posts in a single day, and aggregated those data into per-person, per-day units of observation. We adopted this precedent of “user-days” as a unit of analysis.

Statistical framework

Overview

We used Bayesian hierarchical logistic regression with uninformative priors to determine the strength of individual predictors. Two separate models were trained. The All-data model used all collected data to address Hypothesis 1. The Pre-diagnosis model used all data collected from
healthy participants, but only pre-diagnosis data from depressed participants, to address Hypothesis 2. We also fit an “intercept-only” model, in which all predictors are zero-weighted to simulate a model under a null hypothesis. Bayes factors were used to assess model fit. Details on Bayesian estimation, model optimization and selection, and diagnostic checks are described in the sections that follow.

We also employed a suite of supervised machine learning algorithms to estimate the predictive capacity of our models, including regularized logistic regression, Random Forests, and Support Vector Machines. We report prediction results only from the best-performing algorithm, a 1200-tree Random Forests classifier. Random Forests hyperparameter optimization was achieved using a grid search over likely parameters (see below for details). As an informal benchmark for comparison, we present general practitioners’ unassisted diagnostic accuracy as reported in Mitchell, Vaze, and Rao (MVR) (2009)\textsuperscript{19}.

**Bayesian logistic regression**

A Bayesian framework avoids many of the inferential challenges of frequentist null hypothesis significance testing, including reliance on p-values and confidence intervals, both of which are subject to frequent misuse and misunderstanding (Gigerenzer, 2004; Hubbard & Lindsay, 2008; Morey et al., 2015; Wasserstein & Lazar, 2016). Logistic regression was conducted using the MCMClogit function from the R package MCMCpack (Martin, Quinn, & Park, 2011). We chose “uninformative” priors for all parameters in $\beta$, with function parameters $b_0 = 0, B_0 = 0.0001$. While generally it is preferable to specify Bayesian priors, in this setting our parameters of interest were entirely novel, and so were not informed by prior literature or previous testing.

\textsuperscript{19} Comparing point estimates of accuracy metrics is not a statistically robust means of model comparison. However, we felt it was more meaningful to frame our findings in a realistic context, rather than to benchmark against a naive statistical model that simply predicted the majority class for all observations.
The MCMClogit function employs a Metropolis algorithm to perform Markov Chain Monte Carlo (MCMC) simulations. We generated two MCMC chains of 100,000 iterations with a burn-in of 10,000 and no thinning. The use of thinning for achieving higher-precision estimates from posterior samples is questionable when compared to simply running longer chains (Link & Eaton, 2012). While no best practice has been established for how long an unthinned chain should be, Christensen et al. (2011) advised: “Unless there is severe autocorrelation, e.g., high correlation with, say \[\text{lag}=30\], we don't believe that thinning is worthwhile”. In our MCMC chains, we observed low autocorrelation at a lag of 30, and so felt confident in foregoing thinning. For comparison, we also ran a 100,000-iteration chain, thinned to every 10th iteration, with a burn-in of 5,000. While autocorrelation was noticeably reduced at shorter lags, this chain yielded near-identical parameter estimates from the posterior sampling distribution.

Recall that Bayesian regression coefficients are not assigned p-values or any other significance measures conventional in frequentist null-hypothesis significance testing (NHST). We have provided Highest Posterior Density Intervals (HPDIs) for the highest probability at which the interval excludes zero as a possible coefficient value. For example, if a 99% HPDI is reported, it means that, based on averaged samples from the simulated joint posterior distribution, the coefficient in question has a 99% probability of being non-zero. References to variable “significance” in the Results section relate only to the probability that a variable’s parameter estimate is non-zero, e.g. “Variable X was significant with 99% probability”.

Bayes factors were used to assess model fit. Given two models \(M_a, M_b\) parameterized by parameter vectors \(\theta_a, \theta_b\), and data \(D\), the Bayes factor is computed as the ratio

\[
K = \frac{P_r(D|\theta_a)}{P_r(D|\theta_b)} = \frac{\int P_r(\theta_a|M_a)P_r(D|\theta_a, M_a) \, d\theta_a}{\int P_r(\theta_b|M_b)P_r(D|\theta_b, M_b) \, d\theta_b}
\]
A positive-valued Bayes factor supports model $M_a$ over $M_b$. Jeffreys (1961) established benchmarks for interpreting $K$ in terms of evidence for $M_a$ as the stronger model (see Table 2).

<table>
<thead>
<tr>
<th>$K$</th>
<th>Strength of evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>$&lt; 10^0$</td>
<td>Negative (supports $M_b$)</td>
</tr>
<tr>
<td>$10^0$ to $10^{1/2}$</td>
<td>Barely worth mentioning</td>
</tr>
<tr>
<td>$10^{1/2}$ to $10^1$</td>
<td>Substantial</td>
</tr>
<tr>
<td>$10^1$ to $10^{3/2}$</td>
<td>Strong</td>
</tr>
<tr>
<td>$10^{3/2}$ to $10^2$</td>
<td>Very strong</td>
</tr>
<tr>
<td>$&gt; 10^2$</td>
<td>Decisive</td>
</tr>
</tbody>
</table>

Markov Chain Monte Carlo (MCMC) chains showed good convergence across all estimated parameters on every fitted model. In all models, Gelman-Rubin diagnostics (Gelman & Rubin, 1992) indicated simulation chain convergence, with point estimates of 1.0 for each parameter. Geweke diagnostics (Geweke, 1996) also indicated post-burn-in convergence. Autocorrelation was observed within acceptable levels. Trace, density, and autocorrelation plots for all models are presented in Results.

*Logistic regression assumptions and grouped observations*

The integrity of logistic regression output relies on several assumptions about the data it receives as input. In particular, one assumption is that observations are independent of one another, which is not necessarily the case with our selected units of observation, user-days. A single observation is an aggregate of a day’s worth of posts, for a given user, meaning that all user-days are conceptually grouped by users (i.e. participants). It is reasonable to assume that the posts made by Participant A on Tuesday are more similar to posts made by Participant A on
Monday, than they are similar to posts made by Participant B on any day. As such, all of Participant A’s user-days are likely to share similarities, as are all of Participant B’s user-days, and so on for all participants. This structural variance between observations violates the observational independence assumption, and can result in dramatically undervalued estimates of variance for parameter means.

Within-user covariance between observations can be captured by explicitly defining a user level in the model. This multi-level, or hierarchical, form of logistic regression is a more appropriate way to conduct this part of our analysis (although note that the Random Forests model, which we use to assess true predictive accuracy, carries no assumptions of observational independence). MCMC simulations for a hierarchical logistic regression model\(^{20}\), however, proved unstable - trace plots, Geweke, and Gelman-Rubin metrics all showed considerable divergence, autocorrelation, and failure to stabilize, even after 200,000 iterations per simulation chain. This is likely due to the fact that a hierarchical structure effectively forces the regression algorithm to consider a much smaller sample size, as there are only 166 users for grouping, despite many thousands of user-days, and so becomes limited in its ability to find global (or even local) optima. We chose to report results from a standard, non-hierarchical logistic regression (for which MCMC diagnostics showed stable convergence), with the considerable caveat that the variance on parameter estimates may be larger than reported.

**Machine learning models**

We employed a suite of supervised machine learning algorithms to estimate the predictive capacity of our models. In a supervised learning paradigm, parameter weights are determined by training on a labeled subset of the total available data ("labeled" here means that the response classes are exposed). Fitted models are then used to predict class membership for

\(^{20}\) We used the MCMCglmm module for hierarchical logistic regression (Hadfield, 2010).
each observation in the remaining unlabeled “holdout” data. All of our machine learning classifiers were trained on a randomly-selected 70% of total observations, and tested on the remaining 30%. We employed stratified five-fold cross-validation to optimize hyperparameters, and averaged final model output metrics over five separate randomized runs. Cross-validation and averaging over multiple randomized runs are common techniques in machine learning to reduce the chance of overfitting a model to data which are unrepresentative of the overall sample.

In evaluating binary classification accuracy, a simple proportion of correct classifications (“naive accuracy”) is often inappropriate. In cases where data exhibit a class imbalance, i.e. more healthy than depressed observations (or vice-versa), reporting naive accuracy can be misleading. (A classification accuracy of 95% seems excellent until it is revealed that 95% of the data modeled belong to a single class.) Additionally, naive accuracy scores are opaque to the specific strengths and weaknesses of a binary classifier. Instead, we report precision, recall, specificity, negative predictive value, and F1 scores for fuller context. See Chapter II for an extended discussion on accuracy metrics for classification algorithms.

**Results**

All data collection took place between February 1, 2016 and April 6, 2016. Across both depressed and healthy groups, we collected data from 166 Instagram users, and analyzed 43,950 posted photographs. The mean number of posts per user was 264.76 (SD=396.06). This distribution was skewed by a smaller number of frequent posters, as evidenced by a median value of just 122.5 posts per user.

---

21 The optimization routine performed a grid search over every combination of the recommended values for Random Forests hyperparameters in Thakur (2016).
Among depressed participants, 84 individuals successfully completed participation and provided access to their Instagram data. Imposing the CES-D cutoff reduced the number of viable participants to 71. The mean age for viable participants was 28.8 years (SD=7.09), with a range of 19 to 55 years. Dates of participants’ first depression diagnoses ranged from February 2010 to January 2016, with nearly all diagnosis dates (90.1%) occurring in the period 2013-2015.

Among healthy participants, 95 participants completed participation and provided access to their Instagram data. The mean age for this group was 30.7 years, with a range of 19 to 53 years, and 65.3% of respondents were female. (Gender data were not collected for the depressed sample.)

All-data model data consisted of participants’ entire Instagram posting histories, consisted of 43,950 Instagram posts (24,811 depressed) over 166 individuals (71 depressed). Aggregation by user-days compressed into 24,713 observations (13,230 depressed). Observations from depressed participants accounted for 53.4% of the entire dataset.

Pre-diagnosis model data used only Instagram posts from depressed participants made prior to the date of first depression diagnosis, along with the same full dataset from healthy participants as used in the All-data model. These data consisted of a total of 32,311 posts (13,192 depressed). There were 18,513 aggregated-unit observations in total (7,030 depressed). Observations from depressed participants accounted for 38% of this dataset. See Table 3.
Table 3
Summary statistics for Instagram data collection (N=43,950).

<table>
<thead>
<tr>
<th></th>
<th>Users</th>
<th>Posts</th>
<th>Posts (μ)</th>
<th>Posts (σ)</th>
<th>Posts (median)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>166</td>
<td>43,950</td>
<td>264.76</td>
<td>396.06</td>
<td>122.5</td>
</tr>
<tr>
<td>Depressed</td>
<td>71</td>
<td>24,811</td>
<td>349.45</td>
<td>441.19</td>
<td>196.0</td>
</tr>
<tr>
<td>Healthy</td>
<td>95</td>
<td>19,139</td>
<td>201.46</td>
<td>347.76</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Both All-data and Pre-diagnosis models were decisively superior to a null model ($K_{all} = 157.5; K_{pre} = 149.8$). All-data predictors were significant with 99% probability. Pre-diagnosis and All-data confidence levels were largely identical, with two exceptions: Pre-diagnosis Brightness decreased to 90% confidence, and Pre-diagnosis posting frequency dropped to 30% confidence, suggesting a null predictive value in the latter case.

Increased hue, along with decreased brightness and saturation, predicted target class observations. This means that photos posted by depressed individuals tended to be bluer, darker, and grayer (see Figure 3). The more comments Instagram posts received, the more likely they were posted by depressed participants, but the opposite was true for likes received. In the All-data model, higher posting frequency was also associated with depression. Depressed participants were more likely to post photos with faces, but had a lower average face count per photograph than healthy participants. Finally, depressed participants were less likely to apply Instagram filters to their posted photos. See Tables 4 and 5.
Figure 3. Magnitude and direction of regression coefficients in All-data (N=24,713) and Pre-diagnosis (N=18,513) models. X-axis values represent the adjustment in odds of an observation belonging to the target class, per unit increase of each predictive variable. Odds were generated by exponentiating logistic regression log-odds coefficients.
Table 4
Logistic regression output for All-data model 1 (N=24,713)

<table>
<thead>
<tr>
<th></th>
<th>Depressed µ (σ)</th>
<th>Healthy µ (σ)</th>
<th>Coef µ (σ)</th>
<th>HPD Level</th>
<th>HPD Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>- .485 (.677)</td>
<td>- .021</td>
<td>65%</td>
<td>-1.214</td>
<td>.023</td>
</tr>
<tr>
<td>Hue</td>
<td>.345 (.162)</td>
<td>.338 (.157)</td>
<td>.139 (.703)</td>
<td>95%</td>
<td>.023</td>
</tr>
<tr>
<td>Saturation</td>
<td>.338 (.157)</td>
<td>.347 (.155)</td>
<td>- .038 (.228)</td>
<td>55%</td>
<td>.003</td>
</tr>
<tr>
<td>Brightness</td>
<td>.535 (.138)</td>
<td>.547 (.145)</td>
<td>-1.063 (.500)</td>
<td>95%</td>
<td>-1.902</td>
</tr>
<tr>
<td>Comments</td>
<td>1.077 (2.150)</td>
<td>.992 (2.013)</td>
<td>.028 (.033)</td>
<td>90%</td>
<td>.002</td>
</tr>
<tr>
<td>Likes</td>
<td>16.168 (34.874)</td>
<td>18.939 (34.214)</td>
<td>-.003 (.002)</td>
<td>85%</td>
<td>-.008</td>
</tr>
<tr>
<td>Posts/day</td>
<td>1.875 (1.961)</td>
<td>1.667 (1.775)</td>
<td>.051 (.050)</td>
<td>99%</td>
<td>.030</td>
</tr>
<tr>
<td>Has filter</td>
<td>.829 (1.108)</td>
<td>.871 (1.524)</td>
<td>-.079 (.046)</td>
<td>99%</td>
<td>-.167</td>
</tr>
<tr>
<td>Has face</td>
<td>.769 (1.137)</td>
<td>.615 (.882)</td>
<td>.122 (.052)</td>
<td>95%</td>
<td>.067</td>
</tr>
<tr>
<td>Face count</td>
<td>.631 (.897)</td>
<td>.623 (.984)</td>
<td>-.099 (.091)</td>
<td>90%</td>
<td>-.240</td>
</tr>
</tbody>
</table>

*HPD Level = Highest Posterior Density Level, the probability that a regression coefficient falls within the given HPD Interval. HPD Levels listed are highest probabilities with which it can be claimed that a coefficient’s HPD Interval excludes zero.


Table 5
Logistic regression output for Pre-diagnosis model (N=18,513)

<table>
<thead>
<tr>
<th></th>
<th>Depressed µ (σ)</th>
<th>Healthy µ (σ)</th>
<th>Coef µ (σ)</th>
<th>HPD Levela</th>
<th>HPD Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td></td>
<td></td>
<td>.463 (.093)</td>
<td>99%</td>
<td>-.695 -.211</td>
</tr>
<tr>
<td>Hue</td>
<td>.360 (.166)</td>
<td>.338 (.157)</td>
<td>.802 (.099)</td>
<td>99%</td>
<td>.545 1.054</td>
</tr>
<tr>
<td>Saturation</td>
<td>.348 (.157)</td>
<td>.347 (.155)</td>
<td>-.271 (.104)</td>
<td>99%</td>
<td>-.522 -.002</td>
</tr>
<tr>
<td>Brightness</td>
<td>.534 (.136)</td>
<td>.547 (.145)</td>
<td>-.209 (.118)</td>
<td>90%</td>
<td>-.410 -.026</td>
</tr>
<tr>
<td>Comments</td>
<td>.912 (1.771)</td>
<td>.992 (2.013)</td>
<td>.047 (.010)</td>
<td>99%</td>
<td>.023 .072</td>
</tr>
<tr>
<td>Likes</td>
<td>12.719 (28.912)</td>
<td>18.939 (34.214)</td>
<td>-.008 (.001)</td>
<td>99%</td>
<td>-.010 -.007</td>
</tr>
<tr>
<td>Posts/day</td>
<td>1.877 (1.931)</td>
<td>1.667 (1.775)</td>
<td>-.004 (.015)</td>
<td>30%</td>
<td>-.012 .000</td>
</tr>
<tr>
<td>Has filter</td>
<td>.907 (1.191)</td>
<td>.871 (1.524)</td>
<td>-.071 (.015)</td>
<td>99%</td>
<td>-.108 -.030</td>
</tr>
<tr>
<td>Has face</td>
<td>.743 (1.030)</td>
<td>.615 (.882)</td>
<td>.267 (.029)</td>
<td>99%</td>
<td>.189 .340</td>
</tr>
<tr>
<td>Face count</td>
<td>.57 (.824)</td>
<td>.623 (.984)</td>
<td>-.207 (.024)</td>
<td>99%</td>
<td>-.268 -.144</td>
</tr>
</tbody>
</table>

\(^a\) HPD Level = Highest Posterior Density Level, the probability that a regression coefficient falls within the given HPD Interval. HPD Levels listed are highest probabilities with which it can be claimed that a coefficient’s HPD Interval excludes zero.

A posterior predictive check showed that All-data observations replicated from the joint posterior distribution consistently overestimated the proportion of depressed observations (replicated: 53.5% depressed; original: 30.9%), with a p-value of 1.0\(^22\). Pre-diagnosis observations sampled from the joint posterior distribution slightly underestimated the proportion of depressed observations (replicated: 30.02% depressed; original: 37.97%), with a posterior predictive p-value of 0.039. Gelman et al. (48) suggested that a model with good replication accuracy should generate posterior predictive p-values within the range of 0.05-0.95. Note that

\(^22\) In the context of logistic regression, the posterior predictive p-value assesses the frequency with which samples drawn from the simulated posterior overpredicts reference class membership, compared to reference class prevalence in the original data.
an extreme posterior predictive p-value does not mean that a model is wrong, just that it fails to be “right enough” to render a reasonable replication of its input. All models nevertheless far outperformed a simple null model in the capacity to correctly predict class membership.

A posterior predictive check of the Ratings model showed that sample observations replicated from the joint posterior distribution accurately represented the true proportion of depressed observations (replicated: 44.2% depressed; original: 43.9%), with a posterior predictive p-value of 0.516.

A closer look at filter usage in depressed versus healthy participants provided additional texture. Instagram filters were used differently by target and control groups ($\chi^2_{all} = 907.84, p = 9.17e-164; \chi^2_{pre} = 813.80, p = 2.87e-144$). In particular, depressed participants were less likely than healthy controls to use any filters at all. When depressed participants did employ filters, they most disproportionately favored the “Inkwell” filter, which converts color photographs to black-and-white images. Conversely, healthy participants most disproportionately favored the Valencia filter, which lightens the tint of photos. See Figure 4 for inter-group filter preferences, and Figure 5 for examples of filtered photographs.
Figure 4. Instagram filter usage among depressed and healthy participants. Bars indicate difference between observed and expected usage frequencies, based on a Chi-squared analysis of independence. Blue bars indicate disproportionate use of a filter by depressed compared to healthy participants, orange bars indicate the reverse. All data results are displayed.
Figure 5. Examples of Inkwell and Valencia Instagram filters. Inkwell converts color photos to black-and-white, Valencia lightens tint. Depressed participants most favored Inkwell compared to healthy participants, Healthy participants most favored Valencia compared to depressed participants.

Image credit: fiterfakers.com

Our best All-data machine learning classifier, averaged over five randomized iterations, improved over MVR general practitioner accuracy on most metrics. Compared with MVR results, the All-data model was less conservative (lower specificity) but better able to positively identify target class observations (higher recall). Given 100 observations, our model correctly
identified 70% of all target class cases (n=37), with a relatively low number of false alarms (n=23) and misses (n=17).

Pre-diagnosis predictions showed improvement over the MVR benchmark on precision and specificity. The Pre-diagnosis model found only about a third of actual target class observations, but it was correct most of the time when it did predict a target class label. By comparison, although MVR general practitioners discovered more true cases of depression, they were more likely than not to misdiagnose healthy subjects as depressed. See Table 6.

<table>
<thead>
<tr>
<th>Table 6</th>
<th>Comparison of accuracy metrics for All-data and Pre-diagnosis model predictions.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MVR µ</td>
</tr>
<tr>
<td>Recall</td>
<td>.510</td>
</tr>
<tr>
<td>Specificity</td>
<td>.813</td>
</tr>
<tr>
<td>Precision</td>
<td>.42</td>
</tr>
<tr>
<td>NPV</td>
<td>.858</td>
</tr>
<tr>
<td>F1</td>
<td>.461</td>
</tr>
</tbody>
</table>

Note. General practitioners’ diagnostic accuracy from Mitchell, Vaze, & Rao (2009) (MVR) is included for comparison.

Rater agreement was measured by randomly selecting two raters from each photo, and computing Pearson’s product-moment correlation coefficient from the resulting vectors. To mitigate sampling bias, we ran a five-fold iteration of this process and averaged the resulting coefficients. Rater agreement showed positive correlations across all ratings categories, \( p < 1e-38 \) for all values shown): r\(_{\text{happy}}\) = .39, r\(_{\text{sad}}\) = .19, r\(_{\text{interesting}}\) = .17, r\(_{\text{likable}}\) = .27.

Out of the four predictors used in the human ratings model (happiness, sadness, likability, interestingness), only the sadness and happiness ratings were significant predictors of depression (see Table 7). Depressed participants’ photos were more likely to be sadder and less happy than
those of healthy participants. Ratings assessments generally showed strong patterns of correlation with one another, but exhibited extremely low correlation with computational features. The modest positive correlation of human-rated happiness with the presence and number of faces in a photograph was the only exception to this trend. Correlation matrices for All-data, Pre-diagnosis, and Ratings models are provided below in Tables 8, 9, and 10, respectively.

<table>
<thead>
<tr>
<th></th>
<th>Depressed μ (σ)</th>
<th>Healthy μ (σ)</th>
<th>Coef μ (σ)</th>
<th>HPD Levela</th>
<th>HPD Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-2.374 (.175)</td>
<td>-2.374 (.175)</td>
<td>-2.374 (.175)</td>
<td>20%</td>
<td>-.045 -.004</td>
</tr>
<tr>
<td>Happy</td>
<td>2.300 (1.042)</td>
<td>2.511 (1.109)</td>
<td>-0.193 (.034)</td>
<td>99%</td>
<td>-.279 -.105</td>
</tr>
<tr>
<td>Sad</td>
<td>.840 (.598)</td>
<td>.757 (.614)</td>
<td>.100 (.039)</td>
<td>95%</td>
<td>.024 .176</td>
</tr>
<tr>
<td>Likable</td>
<td>2.393 (.918)</td>
<td>2.514 (.952)</td>
<td>.027 (.050)</td>
<td>35%</td>
<td>.007 .052</td>
</tr>
<tr>
<td>Interesting</td>
<td>2.316 (.816)</td>
<td>2.367 (.859)</td>
<td>.041 (.041)</td>
<td>65%</td>
<td>.003 .080</td>
</tr>
</tbody>
</table>

a HPD Level = Highest Posterior Density Level, the probability that a regression coefficient falls within the given HPD Interval. HPD Levels listed are highest probabilities with which it can be claimed that a coefficient’s HPD Interval excludes zero.
Table 8
Pearson’s product-moment correlation scores for All-data model features.

<table>
<thead>
<tr>
<th></th>
<th>Hue</th>
<th>Satur.</th>
<th>Bright.</th>
<th>Comm.</th>
<th>Likes</th>
<th>Posts</th>
<th>Filter</th>
<th>Face</th>
</tr>
</thead>
<tbody>
<tr>
<td>Saturation</td>
<td>.17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brightness</td>
<td>-.22</td>
<td>-.28</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comments</td>
<td>-.07</td>
<td>-.05</td>
<td>.10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Likes</td>
<td>-.09</td>
<td>-.10</td>
<td>.17</td>
<td>.55</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Posts</td>
<td>.05</td>
<td>.03</td>
<td>-.06</td>
<td>-.02</td>
<td>-.03</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Has filter</td>
<td>.12</td>
<td>.08</td>
<td>-.02</td>
<td>-.05</td>
<td>-.07</td>
<td>.56</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Has face</td>
<td>.03</td>
<td>.04</td>
<td>-.05</td>
<td>-.00</td>
<td>-.02</td>
<td>.69</td>
<td>.34</td>
<td></td>
</tr>
<tr>
<td>Face count</td>
<td>.01</td>
<td>.04</td>
<td>-.03</td>
<td>.02</td>
<td>.00</td>
<td>-.02</td>
<td>-.01</td>
<td>.42</td>
</tr>
</tbody>
</table>
Table 9
Pearson’s product-moment correlation scores for Pre-diagnosis model features.

<table>
<thead>
<tr>
<th></th>
<th>Hue</th>
<th>Satur.</th>
<th>Bright.</th>
<th>Comm.</th>
<th>Likes</th>
<th>Posts</th>
<th>Filter</th>
<th>Face</th>
</tr>
</thead>
<tbody>
<tr>
<td>Saturation</td>
<td>.16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brightness</td>
<td>-.22</td>
<td>-.27</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comments</td>
<td>-.05</td>
<td>-.06</td>
<td>.14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Likes</td>
<td>-.08</td>
<td>-.13</td>
<td>.23</td>
<td>.49</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Posts</td>
<td>.06</td>
<td>.03</td>
<td>-.06</td>
<td>-.04</td>
<td>-.06</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Has filter</td>
<td>.13</td>
<td>.09</td>
<td>-.03</td>
<td>-.04</td>
<td>-.08</td>
<td>.63</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Has face</td>
<td>.04</td>
<td>.05</td>
<td>-.05</td>
<td>-.01</td>
<td>-.05</td>
<td>.66</td>
<td>.39</td>
<td></td>
</tr>
<tr>
<td>Face count</td>
<td>.01</td>
<td>.03</td>
<td>-.02</td>
<td>.03</td>
<td>-.01</td>
<td>-.02</td>
<td>-.02</td>
<td>.45</td>
</tr>
</tbody>
</table>
Table 10  
Pearson’s product-moment correlation scores for Ratings model features (columns) with ratings and computational features (rows)

<table>
<thead>
<tr>
<th></th>
<th>Happy</th>
<th>Sad</th>
<th>Likable</th>
<th>Interest.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sad</td>
<td></td>
<td>-0.41</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Likable</td>
<td>0.79</td>
<td>-0.29</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interesting</td>
<td>0.53</td>
<td>-0.09</td>
<td>0.77</td>
<td></td>
</tr>
<tr>
<td>Hue</td>
<td>0.02</td>
<td>-0.02</td>
<td>-0.01</td>
<td>-0.03</td>
</tr>
<tr>
<td>Saturation</td>
<td>0.02</td>
<td>-0.07</td>
<td>-0.02</td>
<td>-0.04</td>
</tr>
<tr>
<td>Brightness</td>
<td>0.05</td>
<td>-0.04</td>
<td>0.04</td>
<td>0.03</td>
</tr>
<tr>
<td>Posts</td>
<td>-0.02</td>
<td>0.04</td>
<td>-0.01</td>
<td>0.02</td>
</tr>
<tr>
<td>Comments</td>
<td>0.00</td>
<td>0.02</td>
<td>-0.02</td>
<td>-0.03</td>
</tr>
<tr>
<td>Likes</td>
<td>0.04</td>
<td>-0.02</td>
<td>0.05</td>
<td>0.06</td>
</tr>
<tr>
<td>Has filter</td>
<td>0.03</td>
<td>0.00</td>
<td>0.02</td>
<td>0.01</td>
</tr>
<tr>
<td>Has face</td>
<td>0.16</td>
<td>0.05</td>
<td>0.06</td>
<td>0.00</td>
</tr>
<tr>
<td>Face count</td>
<td>0.25</td>
<td>-0.10</td>
<td>0.11</td>
<td>0.02</td>
</tr>
</tbody>
</table>

**Discussion**

The present study employed computational machine learning techniques to screen for depression using photographs posted to Instagram. Our results supported Hypothesis 1, that markers of depression are observable in Instagram user behavior, and Hypothesis 2, that these depressive signals are detectable in posts made even before the date of first diagnosis. Human ratings proved capable of distinguishing between Instagram posts made by depressed and healthy individuals (Hypothesis 3a), but showed little or no correlation with most computational features
(Hypothesis 3b). Our findings establish that visual social media data are amenable to analysis of affect using scalable, computational methods. One avenue for future research might integrate textual analysis of Instagram posts’ comments, captions, and tags. Considering the early success of textual analysis in detecting various health and psychological signals on social media (De Choudhury et al., 2013; Dodds et al., 2011), the modeling of textual and visual features together could well prove superior to either medium on its own.

Our model showed considerable improvement over the ability of unassisted general practitioners to correctly diagnose depression. On average, more than half of general practitioners’ depression diagnoses were false positives (Mitchell, Vaze, & Rao, 2009). By comparison, the majority of both All-data and Pre-diagnosis depression classifications were correct. As false diagnoses are costly for both healthcare programs and individuals, this improvement is noteworthy. Health care providers may be able to improve quality of care and better identify individuals in need of treatment based on the simple, low-cost methods outlined in this report. Given that mental health services are unavailable or underfunded in many countries (Detels, 2009), this computational approach, requiring only patients’ digital consent to share their social media histories, may open avenues to care which are currently difficult or impossible to provide.

On the other hand, our Pre-diagnosis prediction engine was rather conservative, and tended to classify most observations as healthy. There is good reason to believe, however, that the Pre-diagnosis prediction accuracy observed represents a lower bound on performance. Ideally, we would have used the All-data classifier to evaluate the Pre-diagnosis data, as that model was trained on a much larger dataset. The fact that the Pre-diagnosis data was a subset of the full dataset meant that applying the All-data model to Pre-diagnosis observations would have
artificially inflated accuracy, due to information leakage between training and test data. Instead, we trained a new classifier for Pre-diagnosis, using training and test partitions contained within the Pre-diagnosis data, which left the Pre-diagnosis model with considerably fewer data points to train on. As a result, it is likely that Pre-diagnosis accuracy scores understate the technique’s true capacity.

Regarding the strength of specific predictive features, some results match common perceptions regarding the effects of depression on behavior. Photos posted to Instagram by depressed individuals were more likely to be bluer, grayer, and darker, and receive fewer likes. Depressed Instagram users in our sample had an outsized preference for filtering out all color from posted photos, and showed an aversion to artificially lightening photos, compared to non-depressed controls. These results matched well with the literature linking depression and a preference for darker, bluer, and monochromatic colors (Barrick, Taylor, & Correa, 2002; Boyatzis & Varghese, 1994; Carruthers, Morris, Tarrier, & Whorwell, 2010; Hemphill, 1996). Depressed users were more likely to post photos with faces, but they tended to post fewer faces per photo. This finding may be an oblique indicator that depressed users interact in smaller social settings, or at least choose only to share experiences of this sort on social media. This would be in accordance with previous findings that reduced social interactivity is an indicator of depression (American Psychiatric Association, 2000; Bruce & Hoff, 1994; De Choudhury et al., 2013a).

Other, seemingly obvious, relationships failed to emerge. For example, when people rated a photograph as sad, that impression was unrelated to how blue, dark, or gray that photo was. Both “sad” and “blue, dark, and gray” were strong predictors of depression, however, and semantically these descriptions seem like they should match well with one another, as well as
link to depression. These divergences may serve as the basis for a number of future research inquiries into the relationship between depressive behavior and common perceptions of depression.

A general limitation to these findings concerns the non-specific use of the term “depression” in the data collection process. We acknowledge that depression describes a general clinical status, and is frequently comorbid with other conditions. It is possible that a specific diagnostic class is responsible for driving the observed results, and future research should fine-tune questionnaires to acquire specific diagnostic information. Additionally, it is possible that our results are in some way specific to individuals who received clinical diagnoses. Current perspectives on depression treatment indicate that people who are “well-informed and psychologically minded, experience typical symptoms of depression and little stigma, and have confidence in the effectiveness of treatment, few concerns about side effects, adequate social support, and high self-efficacy” seek out mental health services (Epstein et al., 2010). The intersection of these qualities with typical Instagram user demographics suggests caution in making broad inferences, based on our findings.

A notable methodological limitation exists in the structural dependency that exists in the chosen units of observation, user-days. While the primary predictive algorithm, a Random Forests classifier, is not restricted by assumptions of observational independence, the same is not true for logistic regression, which we used to assess the directionality of individual predictors. As discussed in Methods, a superior approach would explicitly model the hierarchical structure of the data, accounting for the shared variance among users’ posts. More data are needed to stabilize the output from a hierarchical model, however, and so this concern remains a topic for improvement in future research.
The choice of user-days as the best unit of observation might also be reconsidered. Just as a lifetime of reported feelings is too long a span, and a single moment too short, in order for a physician to make an accurate diagnosis of a patient’s current mental status, so too our algorithms would have been ill-applied in using an individual’s entire Instagram history, or a single photo, to ascribe a label of depressed or healthy. We chose to follow the precedent of De Choudhury et al. (2013), who addressed the same problem (but for Twitter) by aggregating data into user-days, with the view that a day’s worth of posts provides, on average, a reasonable snapshot into a user’s mental state by which to make meaningful assessments. Even this compromise, however, probably does not accurately reflect the period of time taken into account by health care practitioners when attempting to diagnose depression. Based on informal conversations on this topic, we estimate that anywhere from two weeks’ to several months’ worth of recent history may be explored in a diagnostic interview. One approach for future research would be to conduct a preliminary interview study of health care providers, which asked them how much of a patient’s history they typically inquire about when trying to make a mental health assessment. The resulting sample mean could then serve to define the observational units used in classifying social media data (e.g. user-two-weeks instead of user-days). Note, however, that the longer the period identified as optimal for observational units, the larger the time span of collected data will need to be for each individual participant in order to maintain the same level of statistical power. For example, a participant with three weeks’ worth of posts will have 21 user-days observations, but only a single user-three-weeks observation. As such, any solution to the limitations identified here will necessarily involve finding better methods for large-scale recruitment and data collection.
As these methods provide a tool for inferring personal information about individuals, two points of caution should be considered. First, data privacy and ethical research practices are of particular concern, given recent admissions that individuals’ social media data were experimentally manipulated or exposed without permission (Fiske & Hauser, 2014, Lumb, 2016). It is perhaps reflective of a current general skepticism towards social media research that, of the 509 individuals who began our survey, 221 (43%) refused to share their Instagram data, even after we provided numerous privacy guarantees. Future research should prioritize establishing confidence among experimental participants that their data will remain secure and private. Second, data trends often change over time, leading socio-technical models of this sort to degrade without frequent calibration (Lazer et al., 2014). The findings reported here should not be taken as enduring facts, but rather as promising leads upon which to build and refine subsequent models.

Paired with a commensurate focus on upholding data privacy and ethical analytics, the present work may serve as a blueprint for effective mental health screening in an increasingly digitized society. More generally, these findings support the notion that major changes in individual psychology are transmitted in social media use, and can be identified via computational methods.
Chapter IV. Forecasting the onset and course of mental illness with Twitter data

Social media data provide valuable clues about physical and mental health conditions. This holds true even in cases where social media users are not yet aware that their health has changed. For example, searching for information on certain health symptoms has been shown to provide accurate early-warning indicators for hard-to-detect cancers (Paparrizos et al., 2016). Social media networks have been used to plot the trajectory of disease outbreaks (Christakis & Fowler, 2010; Li & Cardie, 2013; Schmidt, 2012), and to track regional dietary health (Alajajian et al., 2015). In addition to physical ailments, predictive screening methods have successfully identified markers in social media data for a number of mental health issues, including addiction (Moreno et al., 2012), depression (De Choudhury et al., 2013a; De Choudhury et al., 2013b; Katikalapudi et al., 2012; Moreno et al., 2011; Park et al., 2012), Post-Traumatic Stress Disorder (PTSD) (Coppersmith et al., 2014; Nadeem, Horn, & Coppersmith, 2016), and suicidal ideation (De Choudhury et al., 2016). The field of predictive health screening with social media data is still in its infancy, however, and considerable refinements are needed to develop methodologies that can effectively augment health care. In this report, we present a set of improved methods and novel contributions for predicting and tracking depression and PTSD on Twitter.

Depression has emerged as the leading mental health condition of interest among computational social scientists (De Choudhury et al., 2013a; De Choudhury et al., 2013b; Katikalapudi et al., 2012; Moreno et al., 2011; Park et al., 2012), as it is a relatively common mental disorder (Ferrari et al., 2013) and influences a range of behaviors and patterns of communication (American Psychiatric Association, 2000). Underdiagnosis of depression remains a persistent problem; a recent survey of a major metropolitan area found nearly half
(45%) of all cases of major depression were undiagnosed (Gwynn et al., 2008). PTSD, while less common (Stein et al., 2000), is frequently comorbid with major depression (Campbell et al., 2007). Studies have found that PTSD is underdiagnosed or under-treated by a majority of primary-care physicians (Munro, Freeman, & Law, 2004; Taubman-Ben-Ari et al., 2001). The costs of underdiagnosis of these conditions, both to human quality of life and health care systems, are considerable. Computational methods for early screening and diagnosis of depression and PTSD have the potential to make a positive impact on a major public health issue, with minimal associated costs and labor intensity.

**Improvements and novel contributions**

Early efforts to detect depression and PTSD signals in Twitter data have been promising. Park et al. (2012) established that Twitter users suffering from depression tended to post tweets containing more negative emotional sentiment compared to healthy users. De Choudhury et al. (2013b) successfully identified new mothers suffering from postpartum depression, based on changes in Twitter usage and tweet content. In a separate analysis, De Choudhury et al. (2013a) found that depressive signals were observable in tweets made by individuals with Major Depressive Disorder. A small number of studies have attempted to identify PTSD markers in Twitter data (Coppersmith et al., 2014; Nadeem, Horn, & Coppersmith, 2016).

This growing literature has employed progressively more sophisticated methods for making intelligent inferences about Twitter users’ mental health based on their online activity. Despite these advances, we identified a number of methodological issues in recent reports which we have improved upon in the present work. A brief review of these modifications provide motivation for the results that follow.
De Choudhury et al. (2013a) built a predictive model using tweets from depressed individuals posted within a year prior to their self-reported onset of a recent depressive episode. Subjects were included for analysis if they had experienced at least two depressive episodes within that one year period, meaning that the data used to make predictions contained tweets posted after the first onset of depression. The date of first depression diagnosis for each individual was not explicitly accounted for in their model. As a result, model training data may have contained both tweets posted during a previous depressive episode, as well as tweets posted after subjects had already received a formal diagnosis. Both of these possibilities seem especially likely, as depression-related terms such as diagnosis, antidepressants, psychotherapy, and hospitalization were significant predictors in their model, along with the names of specific antidepressant medications (e.g. serotonin, maprotiline, and nefazodone).

We chose to use only tweets posted prior to the date of subjects’ first depression diagnosis, rather than focus on recent depressive episodes, for three reasons. First, self-reported information about depressive symptoms is often inaccurate (Eaton et al., 2000). By contrast, a clinical diagnosis is an explicit event that does not rely on subjective impressions, as may be the case with self-reported onset dates. Second, individuals diagnosed with depression often come to identify with their diagnosis (Cornford et al., 2007; Karp, 1994), and subsequent choices, including how to portray oneself on social media, may be influenced by this identification. It is possible that the predictive signals indicated in De Choudhury et al. (2013a) were not tracking depressive symptoms, per se, but rather identified purposeful communication choices on the part of depressed Twitter users. Third, and most important, if we are able to accurately discriminate between depressed and healthy participants using only tweets posted prior to first diagnosis, this

---

To our knowledge, De Choudhury et al. (2013a) represent the state-of-the-art in depression screening on Twitter, and our own work was strongly informed by their innovative methods. Accordingly, we report model accuracy scores from De Choudhury et al. (2013a) along with our results as a point of comparison.
would support a stronger claim than has been made previously - namely, that Twitter data are capable not only of detecting depression, but can do so before the first diagnosis has been made.

While date of first diagnosis provides a more reliable temporal marker than self-reported onset of symptoms, onset timing is also valuable to researchers and health care professionals looking to better understand depression. This is especially true regarding the onset of an individual’s first depressive episode. Winokur (1976) found that over 50% of depression patients experienced first onset at least 6 months prior to diagnosis. The months during which individuals suffering from depression are undiagnosed and untreated pose a significant health risk. Given that the changes that occur with the first onset of depression may be reflected in social media data, we hypothesized that a computational approach could model the progression of depression without any explicit estimates of onset. Using only the content of participants’ tweets, we generated a time series model which charts the course of illness in depressed individuals, and compared this with healthy participants’ data.

All of the above methodological improvements were also applied to our PTSD analysis, which used a separate cohort of study participants. Extant literature on PTSD detection in Twitter data (Coppersmith et al., 2014; Nadeem, Horn, & Coppersmith, 2016) differ from our analysis in important ways. Previous research used bulk collections of public tweets, and assigned PTSD labels to users based on tweets which mentioned a PTSD diagnosis. By comparison, we communicated directly with participants, and excluded any who could not report a specific date on which they received a professional clinical diagnosis. Our analytical approach incorporated a wide array of metadata features and semantic measures, which were limited (Coppersmith et al., 2014) or missing entirely (Nadeem, Horn, & Coppersmith, 2016) from earlier research. Most importantly, previous research focused only on differentiating PTSD users
from healthy users, without any consideration of timing with respect to the dates of traumatic events or diagnoses. Our models focused specifically on identifying predictive markers of PTSD prior to diagnosis date, as well as tracking the course of this disorder over time.

**Comparison with trained healthcare professionals**

Mitchell, Vaze, and Rao (2009) evaluated general practitioners’ abilities to correctly diagnose depression in their patients, without assistance from scales, questionnaires, or other measurement instruments. Out of 50,371 patient outcomes culled from 118 studies, 21.9% of patients were actually depressed. General practitioners were able to correctly rule out depression in 81% of non-depressed patients, but only correctly diagnosed depressed patients 42% of the time. Taubman-Ben-Ari et al. (2001) tested primary-care physicians’ abilities to detect PTSD. PTSD prevalence was 7.5% for men and 10.5% for women in the observed sample (N=683). Physicians correctly identified 2.5% of PTSD cases, and out of all PTSD diagnoses made, only 43% were accurate. We refer to general practitioner accuracy rates (22,27) as an informal benchmark for the quality of our computational model of depression.

**Method**

This study was reviewed and approved by the Harvard University Institutional Review Board, approval #15-2529 and by the University of Vermont Institutional Review Board, approval #CHRMS-16-135. All study participants were informed of and acknowledged all study goals, expectations, and procedures, including data privacy procedures, prior to any data collection. Surveys were built using the Qualtrics survey platform ([www.qualtrics.com](http://www.qualtrics.com)). Analyses were conducted using the Python (v2.7) programming language ([Python Software Foundation, 2016](http://dev.twitter.com)). Social media data collection apps were written in Python, using the Twitter developer’s Application Programming Interface (API) ([dev.twitter.com](http://dev.twitter.com)).
Data Collection

Participants were recruited using Amazon’s Mechanical Turk (MTurk) crowdwork platform, and we collected user data from both the survey on MTurk and participants’ Twitter history. Recruitment and data collection procedures were identical for depression and PTSD samples, with the exception of the condition-specific questionnaire used for screening. Separate surveys were created for affected and healthy samples. In the affected sample surveys, participants were invited to complete a questionnaire that involved passing a series of inclusion criteria, responding to a standardized clinical assessment survey, answering questions related to demographics and mental health history, and sharing social media history. We used the CES-D (Center for Epidemiologic Studies Depression Scale) questionnaire to screen participant depression levels (Radloff, 1977). CES-D assessment quality has been demonstrated as on-par with other depression inventories, including the Beck Depression Inventory and the Kellner Symptom Questionnaire (Fountoulakis et al., 2007; Zich et al., 1990). The Trauma Screening Questionnaire (TSQ) was used to screen for PTSD (Brewin et al., 2002). A comparison cohort of healthy participants were screened to ensure no history of depression or PTSD, respectively, and for active Twitter use.

Qualified participants were asked to share their Twitter usernames and history. An app embedded in the survey allowed participants to securely log into their Twitter accounts and agree to share their data. Upon securing consent, we made a one-time collection of participants’ Twitter posting history. In total we collected 279,951 tweets from 204 Twitter users for the depression analysis, and 243,775 tweets from 174 Twitter users for the PTSD analysis. Details on participant data protection measures are outlined below.

24 The term “affected” here refers to participants affected by either of the mental health conditions of interest: depression or PTSD.
Inclusion criteria

The surveys for affected samples collected age data from participants, and asked qualified participants questions related to their first clinical diagnosis of either depression or PTSD, as well as questions about social media usage at the time of diagnosis. These questions were given in addition to the CES-D or TSQ scales. The purpose of these questions was to determine:

- The date of first clinical diagnosis of the condition,
- Whether or not the individual suspected having the condition before diagnosis, and,
- If so, the number of days prior to diagnosis that this suspicion began

In the case that participants could not recall exact dates, they were instructed to approximate the actual date.

The survey for healthy participants collected age and gender data from participants. It also asked four questions regarding personal health history, which were used as inclusion criteria for this and three other studies. These questions were as follows:

- Have you ever been pregnant?
- Have you ever been clinically diagnosed with depression?
- Have you ever been clinically diagnosed with Post-Traumatic Stress Disorder?
- Have you ever been diagnosed with cancer?

Participants’ responses to these questions were not used in analysis, and only served to include qualified respondents in each of the various studies, including the depression- and PTSD-related studies reported here.

Participant safety and privacy
This study design raised two important issues regarding ethical research practices, as it concerned both individuals with mental illness and potentially personally identifiable information. We were unable to guarantee strict anonymity to participants, given that usernames and personal information posted to Twitter are often inherently specific to participants’ identities (such as usernames and tweets containing real names). As we potentially had the capacity to link study participants’ identities to sensitive health information, study participants were informed of the risks of being personally identified from their social media data. Participants were assured that no personal identifiers, including usernames, would ever be made public or published in any format. We used Turk Prime, an interface for conducting MTurk studies, to mask participants’ MTurk worker IDs from our records. We made it clear that any links between social media data and private personal health data would be available only to our team of researchers, and participants were able to request to have their data removed at any time.

**Improving data quality**

In an effort to minimize noisy and unreliable data, we applied several quality assurance measures in our data collection process. MTurk workers who have completed at least 100 tasks, with a minimum 95% approval rating, have been found to provide reliable, valid survey responses (Peer et al., 2013). We restricted survey visibility only to workers with these qualifications. Survey access was also restricted to U.S. IP addresses, as MTurk data collected from outside the United States are generally of poorer quality (Litman et al., 2014). All participants were only permitted to take the survey once.

We excluded participants with a total of fewer than five Twitter posts. We also excluded participants with CES-D scores of 21 or lower (depression), or TSQ scores of 5 or lower (PTSD). Studies have indicated that a CES-D score of 22 represents an optimal cutoff for identifying
clinically relevant depression (Cuijpers et al., 2007; Haringsma et al., 2004); an equivalent TSQ cutoff of 6 has been found to be optimal in the case of PTSD (Brewin et al., 2002).

**Summary statistics**

All data collection took place between February 1, 2016 and June 10, 2016. Across both depressed and healthy groups, we collected data from 204 Twitter users, totaling 279,951 tweets. The mean number of posts per user was 1372.71 (SD=1281.74). This distribution was skewed by a smaller number of frequent posters, as evidenced by a median value of just 861 posts per user. See Table 11 for summary statistics.

<table>
<thead>
<tr>
<th>Table 11</th>
<th>Summary statistics for depression and PTSD tweet collection (N\textsubscript{depr}=279,951, N\textsubscript{ptsd}=243,775)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depression Users</td>
<td>Posts</td>
</tr>
<tr>
<td>Total</td>
<td>204</td>
</tr>
<tr>
<td>Depressed</td>
<td>105</td>
</tr>
<tr>
<td>Healthy</td>
<td>99</td>
</tr>
<tr>
<td>PTSD Users</td>
<td>Posts</td>
</tr>
<tr>
<td>Total</td>
<td>174</td>
</tr>
<tr>
<td>Depressed</td>
<td>63</td>
</tr>
<tr>
<td>Healthy</td>
<td>111</td>
</tr>
</tbody>
</table>

In the depressed group, 147 crowdworkers successfully completed participation and provided access to their Twitter data. Imposing the CES-D cutoff reduced the number of viable participants to 105. The mean age for viable participants was 30.3 years (SD=8.34), with a range of 18 to 64 years. Dates of participants’ first depression diagnoses ranged from March 2010 to February 2016, with nearly all diagnosis dates (92%) occurring in the period 2013-2015. In the

---

\textsuperscript{25} This number includes up to 3,200 tweets from each participant’s Twitter history. The analyses in this report focus only on tweets from depressed users created before the date of first depression diagnosis.
healthy group, 99 participants completed participation and provided access to their Twitter data. The mean age for this group was 33.9 years, with a range of 19 to 63 years, and 42% of respondents were female. (Gender data were not collected for affected sample surveys.)

For the PTSD analysis, we collected data from 174 Twitter users, totaling 243,775. The mean number of posts per user was 1372.71 (SD=1281.74). This distribution was skewed by a smaller number of frequent posters, as evidenced by a median value of just 862 posts per user.

In the PTSD sample, 73 crowdworkers successfully completed participation and provided access to their Twitter data. Imposing the TSQ cutoff reduced the number of viable participants to 63. The mean age for viable participants was 30.64 years (SD=7.57), with a range of 21 to 54 years. Dates of participants’ first PTSD diagnoses ranged from April 2010 to December 2015, with nearly all diagnosis dates (94%) occurring in the period 2013-2015. In the healthy group, 111 participants completed participation and provided access to their Twitter data. The mean age for this group was 33.25 years, with a range of 19 to 63 years, and 51% of respondents were female.

**Feature extraction**

We extracted several categories of predictors from the Twitter posts collected. Predictor selection for both depression and PTSD was based on prior machine learning models of depression in Twitter data (De Choudhury et al., 2013a; De Choudhury et al., 2013b), as the two conditions’ high comorbidity rates suggest their predictive signals may exhibit considerable overlap (Campbell et al., 2007). Depressed Twitter users have been observed to tweet less frequently than non-depressed users (De Choudhury et al., 2013a), so we used total tweets per user, per day, as a measure of user activity. Tweet metadata was analyzed to assess average word
count per tweet\textsuperscript{26}, whether or not the tweet was a retweet, and whether or not the tweet was a reply to someone else’s tweet. The labMT, LIWC 2007, and ANEW unigram sentiment instruments were used to quantify the happiness of tweet language (Bradley & Lang, 1999; Dodds et al., 2011; Pennebaker et al., 2015; Reagan et al., 2016). The use of labMT, which has shown strong prior performance in analyzing happiness on Twitter (Cody et al., 2015; Frank et al., 2013), is novel with respect to depression screening; ANEW and LIWC have been successfully applied in previous studies on depression and Twitter (Coppersmith et al., 2014; De Choudhury et al., 2013a; De Choudhury et al., 2013b). LIWC was also used to compile frequency counts of various parts of speech (e.g., pronouns, verbs, adjectives) and semantic categories (e.g., food words, familial terms, profanity) as additional predictors (Pennebaker et al., 2015).

**Units of observation**

Determining the best time span for analysis raises a difficult question: When and for how long does mental illness occur? Receiving a clinical diagnosis of depression or PTSD does not imply that an individual remains in a persistent state of illness, and so to conduct analysis with an individual’s entire posting history as a single unit of observation is a dubious proposition. At the other extreme, to take one tweet as a unit of observation runs the risk of being too granular.

DeChoudhury et al. (2013b) looked at all of a given user’s tweets in a single day, and aggregated those data into per-person, per-day units of observation. In this report we have followed the convention of aggregated “user-days” as a primary unit of analysis, rather than try to categorize a person’s entire history, or analyze each individual tweet\textsuperscript{27}. In our own previous

\textsuperscript{26} A word is defined as a set of characters surrounded by whitespace.

\textsuperscript{27} Occasionally, when reporting results we refer to observations or tweets as “depressed”, e.g. “depressed tweets received fewer likes”. It would be more correct to use the phrase “tweet data from depressed participants, aggregated
research, however, we have found that many Twitter users do not generate enough daily content to make for robust unigram sentiment analysis (Bliss et al., 2012). For completeness, we conducted analyses using both daily and weekly units of observation\(^\text{28}\). Both analyses yielded predictive models of similar strengths, with the weekly model showing a slight, but consistent, edge in performance. We report accuracy metrics from both analyses, but restrict other results to the daily-unit analysis to allow for more direct comparison with previous research.

**Statistical framework**

*Machine learning models*

We trained supervised machine learning classifiers to discriminate between affected and healthy sample members’ observations. Classifiers were trained on a randomly-selected 70% of total observations, and tested on the remaining 30%. Out of several candidate algorithms, a 1200-tree Random Forests classifier demonstrated best performance. Stratified five-fold cross-validation was used to optimize Random Forests hyperparameters, and final accuracy scores were averaged over five separate randomized runs. Precision, recall, specificity, negative predictive value, and F1 accuracy scores are reported, and general practitioners’ unassisted diagnostic accuracy rates as reported in Mitchell, Vaze, and Rao (2009) (MVR) and Taubman-Ben-Ari et al. (2001) (TBA) are used as informal benchmarks for depression and PTSD, respectively\(^\text{29}\).

\(^{28}\) We also considered hourly units of observation, as De Choudhury et al. (2013b) found substantial differences in diurnal patterns between depressed and healthy subjects. In our sample, however, exploratory analysis showed no differences in hourly posting trends between these groups, and we did not conduct analyses using hourly units.

\(^{29}\) Comparing point estimates of accuracy metrics is not a statistically robust means of model comparison, in addition to the fact that our results are drawn from different samples, using different observational units, than our chosen comparisons. However, we felt it was more meaningful to frame our findings in a realistic context, rather than to benchmark against a naive statistical model that simply predicted the majority class for all observations.
**Time series analysis**

Predictive screening methods use indirect indicators, such as language use on social media, to infer health status. We are not actually interested in the average word count of depressed individuals’ tweets, for example, but rather we hope that this measure will allow us some access to the underlying variable we truly care about: depression. Accordingly, state-space models, which use observable data to estimate the status of a latent, or hidden, variable over time, may provide useful insights. We trained a two-state Hidden Markov Model (HMM) to detect differential changes between affected and healthy groups over time.\(^3^0\)

The use of HMM presents an interpretability challenge: how to know whether resulting latent states have any relationship to the clinical condition of interest? Consider the case of depression: Finding evidence that HMM had, in fact, recovered two states from our data that closely resembled the depressed and healthy classes was prerequisite to making any inferences based on HMM output. We addressed this issue by comparing HMM output with mean differences between depressed and healthy observations in the raw data. If the direction of the differences between HMM mean parameter estimates generally agree with the true differences in the data, this provides evidence that the two sample groups in our data (depressed and healthy) are well-characterized by HMM latent states. For example, if depressed observations contained more sad words on average than healthy observations (variable name: “LIWC\_sad”), then the HMM state with the higher LIWC\_sad estimate is more likely to be the depressed one, given that HMM does track depression (i.e. the latent states generated by HMM map onto “depressed” and “healthy”). If, on the other hand, HMM-generated states are weakly or not at all related to

\(^3^0\) During exploratory data analysis, we found that a general mixture model (GMM), optimized to minimize the Bayesian Information Criterion, found a six-state model was most appropriate. We chose to exercise some design prerogative in enforcing a two-state model, as we were interested mainly in dichotomous outcomes. It is possible that the GMM results indicate the presence of degrees of depression severity in the data, and this concept could serve as a fruitful grounds for future research.
depression, there should be no clear alignment between HMM means and means in the raw data. The same procedure was applied when fitting an HMM to the PTSD data.

**Word shift graphs**

Machine learning algorithms provide powerful predictive capability, but most algorithms offer little in the way of context and interpretation. Word shift graphs use the labMT happiness scores, the most important predictor for both depression and PTSD analyses, to show qualitatively how inter-group differences may be driven by the usage of specific words in tweets (Dodds et al., 2011). We present word shift graphs comparing the way tweet language adjusted happiness scores in affected and healthy samples. The visualization ranks words by their contribution to the happiness difference between the two groups (for more explanation of word shift rankings, see Dodds et al., 2015; Storylab, 2014). Word shifts are generated from a different statistical method than the machine learning algorithm we used to make predictions, and so should be treated as exploratory analyses independent of our main findings.

**Results**

For the depression study, we analyzed 74,990 daily observations (23,541 depressed) from 204 individuals (105 depressed). For the PTSD study, we analyzed 54,197 daily observations (13,008 PTSD) from 174 individuals (63 PTSD). Observations from affected sample members accounted for 31.4% and 24% of the entire data sets for depression and PTSD, respectively.

**Machine learning classifier**

Results are reported for both daily and weekly units of observation (see Table 12 and Figures 6 and 7). Our best depression classifier, averaged over cross-validation iterations, improved over both Mitchell et al. (2009) and De Choudhury et al. (2013b) on several metrics. Our depression model’s precision rate was considerably higher, with just over 1 false positive for
every 10 depression diagnoses. By comparison, general practitioners from Mitchell et al. (2009) incorrectly diagnosed patients as having depression in more than half of all diagnoses.

Our best PTSD classifier improved considerably over the primary-care physicians from Taubman-Ben-Ari et al. (2001) (TBA). Whereas more than half of all PTSD diagnoses made by TBA physicians were incorrect, our model was correct in roughly 9 out of every 10 (88.2%) of its PTSD predictions. Model recall rate was strong, with 68.3% discovery of actual PTSD sample observations.

The labMT happiness score was the strongest predictor of both depression and PTSD. Notably, average labMT average happiness over user days showed only modest correlation with ANEW (r_{depr}=.37, r_{ptsd}=.36) and LIWC (r_{depr}=.36, r_{ptsd}=.37), suggesting that labMT identifies relevant prediction signals not fully captured by other sentiment instruments. The additional benefit offered by labMT in this context may be a reflection of its inclusion of the 5000 most frequently used words on Twitter, including slang (Dodds et al., 2011). The second most important variable was word count, which represented the average number of words per tweet. Sentiment-related variables from ANEW and LIWC accounted for most of the remaining top predictors (see Figure 6)^31.

---

^31 As with most decision-tree classifiers, the Random Forests algorithm provides information on the relevance, but not the directionality, of predictors. In other words, we can know how important a variable was to the algorithm, but not if it was positively or negatively associated with the response variable. Word shift graphs, reported below, offer some indication of directionality but are computed differently than Random Forests and should not be used to directly interpret Random Forests output.
Table 12
Classification accuracy metrics for daily and weekly models ($N_{\text{depr}}=74,990$, $N_{\text{ptsd}}=54,197$).

<table>
<thead>
<tr>
<th></th>
<th>Depression</th>
<th>PTSD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MVR $\mu$</td>
<td>DC $\mu$</td>
</tr>
<tr>
<td>Recall</td>
<td>.510</td>
<td>.614</td>
</tr>
<tr>
<td>Specificity</td>
<td>.813</td>
<td>N/A</td>
</tr>
<tr>
<td>Precision</td>
<td>.42</td>
<td>.742</td>
</tr>
<tr>
<td>NPV</td>
<td>.858</td>
<td>N/A</td>
</tr>
<tr>
<td>F1</td>
<td>.461</td>
<td>.672</td>
</tr>
<tr>
<td></td>
<td>TBA $\mu$</td>
<td>NHC $\mu$</td>
</tr>
<tr>
<td>Recall</td>
<td>.249</td>
<td>.82</td>
</tr>
<tr>
<td>Specificity</td>
<td>.979</td>
<td>N/A</td>
</tr>
<tr>
<td>Precision</td>
<td>.429</td>
<td>.86</td>
</tr>
<tr>
<td>NPV</td>
<td>.602</td>
<td>N/A</td>
</tr>
<tr>
<td>F1</td>
<td>.315</td>
<td>.84</td>
</tr>
</tbody>
</table>

Note. Accuracy scores from Mitchell et al. (2009) (MVR), De Choudhury et al. (2013b) (DC), Taubman-Ben-Ari et al. (2001) (TBA), and Nadeem, Horn, & Coppersmith (2016) (NHC) are included for comparison to depression (MVR, DC) and PTSD (TBA, NHC) results. Table cells marked N/A indicate unavailable metrics from previous studies.
Figure 6. ROC curve and top predictors for Random Forests algorithm, for depression and PTSD samples ($N_{\text{depr}}=74,990$, $N_{\text{ptsd}}=54,197$). Predictor names ending in “_happy” are happiness measures; LIWC predictors (Pennebaker et al., 2015) refer to the occurrence of semantic categories (e.g. LIWC_ingest refers to food and eating words, LIWC_swear refers to profanity).
Figure 7. ROC curve and top predictors for Random Forests algorithm using weekly units of observation, for depression and PTSD samples ($N_{depr}=29,328$, $N_{ptsd}=12,676$). Predictor names ending in “_happy” are happiness measures; LIWC predictors (36) refer to the occurrence of semantic categories (e.g. LIWC_ingest refers to food and eating words, LIWC_swear refers to profanity).

**Time series analysis**

A Hidden Markov Model simulated affected and healthy states. HMM states were determined to accurately track with affected and healthy groups by comparing differences in mean parameter estimates between the model fit and original data. Across all 40 predictors, HMM means were in agreement with true means for the depression sample in 38 cases (95% agreement), and were in 100% agreement with true means for the PTSD sample. This evidence
strongly suggested that the two states identified by HMM were closely aligned with the affected and healthy classes in our data, and we have reported HMM results based on this assumption.

See Figures 9 and 10.
Figure 9. Hidden Markov Model showing probability of depression (N=74,990). X-axis represents days from diagnosis. Healthy data are plotted from a consecutive time span of equivalent length. Trend lines represent cubic polynomial regression fits with 95% CI bands, points are aggregations of 14 day periods, with error bars indicating 95% CI on central tendency of daily values.

Figure 10. Hidden Markov Model showing probability of PTSD (N=54,197). X-axis represents days from trauma event. Healthy data are plotted from a consecutive time span of equivalent length. The purple vertical line indicates mean number of days to PTSD diagnosis, post-trauma, and the purple shaded region shows the average period between trauma and diagnosis. Trend lines represent cubic polynomial regression fits with 95% CI bands, points are aggregations of 30 day periods, with error bars indicating 95% CI on central tendency of daily values.
Depressed individuals showed a slightly higher probability of depression even at nine
months prior to diagnosis, and gradually diverged from healthy data points, becoming
pronounced well before diagnosis. Healthy individuals showed a steady, lower probability of
depression, which did not change noticeably over an 18-month period. By three months prior to
diagnosis, depressed subjects showed a marked rise in probability of being in a depressed state,
whereas healthy individuals showed little or no change over the same time period. Post-
diagnosis, probability of depression began to decrease after 3-4 months (90-120 days). This
trajectory matches closely with average improvement time frames observed in therapeutic
programs (Schulberg et al., 1998). Given that HMM constructed latent states from unlabeled
data, it is striking that HMM not only reconstructed the division between depressed and healthy
groups, but also generated a plausible timeline for depression onset and recovery. Similarly,
tweets from individuals with PTSD deviated from healthy tweets within weeks after the date of
the traumatic event that caused PTSD (indicated by the orange line in Figure 10), and well over a
year before the average time to diagnosis (the mean time period from trauma to diagnosis was
586 days). A decrease in PTSD probability can be observed occurring shortly after diagnosis,
indicating possible improvement due to treatment.

**Word shift graphs**

We averaged labMT happiness scores across observations in each class, after the removal
of common neutral words and re-tweeted promotional material (Dodds et al., 2011)\(^{32,33}\). We
observed that tweets authored by the depressed class were sadder (h\(_{avg}\) = 6.01) than the healthy

\(^{32}\) Words were removed with labMT happiness scores between 4 and 6, on a 1-9 scale. This includes many common
parts of speech, including articles and pronouns, which contribute little to understanding inter-group differences in
valenced language.

\(^{33}\) Some of the positive language observed more frequently among healthy individuals came from re-tweets of
promotional or other advertising material (e.g., “win”, “free”, “gift”). We removed obvious promotional retweets
when generating word shift graphs, as their removal did not significantly change mean tweet-happiness differences
between groups, and the resulting graphs gave better impressions of what participants personally tweeted about.
class ($h_{avg} = 6.15$). In Figures 12 and 13, we rank order individual words with respect to their contribution to this observed difference, and display the top contributing words.

The dominant contributor to the difference between depressed and healthy classes was an increase in usage of negative words by the depressed class, including “don’t”, “no”, “not”, “murder”, “death”, “never”, and “sad”. The second largest contributor was a decrease in positive language by the depressed class, relative to the healthy class, including fewer appearances of “photo”, “happy”, “love”, and “fun”. The increased usage of negatively-valenced language by depressed individuals is congruent with previous research (Rude, Gortner, & Pennebaker, 2004).

<table>
<thead>
<tr>
<th>#</th>
<th>Word</th>
<th>+/-</th>
<th>↑/↓</th>
<th>% Cont.</th>
<th>% Cont.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>photo</td>
<td>+</td>
<td>↓</td>
<td>-6.87%</td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td>don’t</td>
<td>-</td>
<td>↑</td>
<td>-4.61%</td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td>no</td>
<td>-</td>
<td>↑</td>
<td>-4.13%</td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td>not</td>
<td>-</td>
<td>↑</td>
<td>-3.72%</td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td>murder</td>
<td>-</td>
<td>↑</td>
<td>-3.36%</td>
<td></td>
</tr>
<tr>
<td>6.</td>
<td>like</td>
<td>+</td>
<td>↑</td>
<td>2.64%</td>
<td></td>
</tr>
<tr>
<td>7.</td>
<td>beach</td>
<td>+</td>
<td>↓</td>
<td>-2.56%</td>
<td></td>
</tr>
<tr>
<td>8.</td>
<td>disaster</td>
<td>-</td>
<td>↓</td>
<td>2.51%</td>
<td></td>
</tr>
<tr>
<td>9.</td>
<td>death</td>
<td>-</td>
<td>↑</td>
<td>-2.37%</td>
<td></td>
</tr>
<tr>
<td>10.</td>
<td>prison</td>
<td>-</td>
<td>↑</td>
<td>-2.27%</td>
<td></td>
</tr>
<tr>
<td>11.</td>
<td>me</td>
<td>+</td>
<td>↑</td>
<td>2.17%</td>
<td></td>
</tr>
<tr>
<td>12.</td>
<td>never</td>
<td>-</td>
<td>↑</td>
<td>-2.12%</td>
<td></td>
</tr>
<tr>
<td>13.</td>
<td>happy</td>
<td>+</td>
<td>↓</td>
<td>-2.03%</td>
<td></td>
</tr>
<tr>
<td>14.</td>
<td>sad</td>
<td>-</td>
<td>↑</td>
<td>-1.91%</td>
<td></td>
</tr>
<tr>
<td>15.</td>
<td>shit</td>
<td>-</td>
<td>↑</td>
<td>-1.82%</td>
<td></td>
</tr>
</tbody>
</table>

Figure 11. Depression word-shift graph revealing contributions to difference in Twitter happiness observed between depressed and healthy participants. In column 3, (-) indicates a relatively negative word, and (+) indicates a relatively positive word, both with respect to the average happiness of all healthy tweets. An up (down) arrow indicates that word was used more (less) by the depressed class. Words on the left (right) contribute to a decrease (increase) in happiness in the depressed class.
Figure 12. PTSD word-shift graph revealing contributions to Twitter happiness observed for PTSD (6.10) and healthy (6.10) participants. In column 3, (-) indicates a relatively negative word, and (+) indicates a relatively positive word, both with respect to the average happiness of all healthy tweets. An up (down) arrow indicates that word was used more (less) by the PTSD class. Words on the left (right) contribute to a decrease (increase) in happiness in the PTSD class. In column 5, % contribution is calculated with respect to the overall average happiness difference between PTSD and healthy participants, which was quite small.

### Discussion

The aim of the present study was to identify predictive markers of depression and PTSD based on users’ Twitter data using computational methods. Our findings strongly support the claim that computational methods can effectively screen Twitter data for indicators of depression and PTSD. Our method identified these mental health conditions earlier and more accurately than the typical performance of trained health professionals, and was more precise than previous computational approaches. Our state-space models portrayed a timeline for depression which is impressively realistic, given that it was generated analyzing only the text of 140-character messages. In addition, HMM identified a rise in probability of PTSD within six months, post-trauma, compared to the average 19 month delay between trauma event and diagnosis.
experienced by the individuals in our sample. Word shifts provided context for the specific differences in language that shifted happiness scores between samples. These advances make improvements on existing predictive screening technology, as well as contribute novel methods for the identification and tracking of mental illness.

The HMM depression timeline is an intriguing finding, and should be treated with both optimism and caution. HMM assigned each data point a probability of belonging to two latent state-spaces while “blind” to our actual states of interest, as affected/healthy labels were removed from HMM training data. Considering that the model could have used criteria completely unrelated to mental health to delineate between the two latent classes, it is noteworthy that the resulting states’ mean estimates for each variable in both PTSD and depression analyses closely resembled the mean estimates for affected and healthy participants’ data, respectively. This adds support for the claim that affected-condition and healthy Twitter users’ data are objectively different, in addition to providing justification for the use of HMM assignments as indicators of depression/PTSD signals at a given point in time. Despite this evidence in favor of applying HMM to analyze mental health trajectories, HMM is an unsupervised learning procedure and so conclusive HMM-based inferences should be approached cautiously and with close attention on validation procedures. The diverging trajectories observed in our HMM time series suggest that, with careful attention to model validity, state-space modeling may be used to identify and track the onset of certain mental illnesses over time, using only Twitter data.

The labMT happiness measure proved to be the most important predictor in our model, and was considerably stronger than ANEW or LIWC happiness indicators. This is in line with a series of previous findings, which have found labMT measures to be a superior for tracking happiness in Twitter data (Reagan et al., 2016), and suggests that future research in this field
should incorporate this instrument for more accurate measurements. That average tweet word count was the second most important predictor is intriguing, especially as increases in word count were positively associated with depression and PTSD. If anything, depression is often characterized by reduced communication (American Psychiatric Association, 2000), although word count is distinct from posting frequency, which was not a significant predictor in our models. The current depression and PTSD literatures are largely devoid of studies relating verbosity to these conditions, and so this finding may motivate new inquiries into behavioral traits of mental health disorders as observed on social media.

From a practical standpoint, our model showed considerable improvement over the ability of unassisted general practitioners to correctly diagnose depression and PTSD. Despite the imprecise nature of this comparison, given the paucity of data currently available to serve as benchmarks for the type of analysis performed in the present study, our model’s relative success seems encouraging. Health care providers may be able to improve quality of care and better identify individuals in need of treatment based on the simple, low-cost methods outlined in this report. Especially given that mental health services are unavailable or underfunded in many countries (Detels, 2009), this computational approach, which only requires patients’ digital consent to share their social media history, may open avenues to care which are currently difficult or impossible to provide. Future investigations would use the same participant pool to collect both health professionals’ assessments, as well as computational models of participant social media behavior, to allow for more precise comparison.

The present findings may be limited by the non-specific use of the term “depression” in participant surveys. While earlier research identified depression predictors in Twitter data for Major Depressive Disorder and postpartum depression (De Choudhury et al., 2013a; De
Choudhury et al., 2013b), we used a more general category in our recruitment and data collection to build a predictive model capable of screening for common depressive signals. We acknowledge that depression diagnoses exist across a clinical spectrum. It is possible that participants with a specific type of depression were responsible for the observed results. Future research might examine other specific depression classes, including manic depression and dysthymia, to determine whether predictive screening models should be segmented per diagnosis type.

It is also possible that inferences from our results are limited specifically to Twitter users who have been diagnosed with depression or PTSD, and who are willing to share their social media history with researchers. Current literature on depression treatment suggests that people who seek out mental health services are usually “well-informed and psychologically minded, experience typical symptoms of depression and little stigma, and have confidence in the effectiveness of treatment, few concerns about side effects, adequate social support, and high self-efficacy” (Epstein et al., 2010). Since it is possible only a subset of Twitter users will fit this description, we recommend making conservative inferences about depression, as well as PTSD, based on our findings.

Considering the frequent comorbidity of depression and PTSD (Campbell et al., 2007), together with the similarity in predictor importance observed across our analyses of these two conditions, the signals driving our predictive models may share considerable overlap. While our results do not offer strict segregation between these two conditions, this gives little cause for concern when considered from a mental health screening perspective. If the desired outcome is to identify individuals who may be in need of mental health services, whether an individual is flagged for evaluation for depression with possible associated PTSD, or vice-versa, becomes an
academic distinction. If anything, the issue of comorbidity may serve as a useful reminder that
this computational method should not be regarded as a standalone diagnostic tool, but rather as a
technology for early identification of potential mental health issues.

As the methods employed in the present study aim to infer health related information
about individuals, some additional cautionary considerations are in order. Data privacy and
ethical research practices are of particular concern, given recent admissions that individuals’
Facebook and dating profile data were experimentally manipulated or exposed without
permission (Fiske & Hauser, 2014; Lumb, 2016). Indeed, we observed a response rate reflecting
a seemingly reluctant population. Of the 2,261 individuals who began our survey, 790 (35%)
refused to share their Twitter username and history, even after we identified ourselves as an
“academic, not-for-profit research team” and provided the above-mentioned guarantees about
data privacy. Future research should prioritize establishing confidence among experimental
participants that their data will remain secure and private. Complicating efforts to build socio-
technical tools such as the models presented in this study, data trends often change over time,
degrading model performance without frequent calibration (Lazer et al., 2014). As such, our
results should be considered a methodological proof-of-concept upon which to build and refine
subsequent models.

This report provides an outline for an accessible, accurate, and inexpensive means of
improving depression and PTSD screening, especially in contexts where in-person assessments
are difficult or costly. In concert with robust data privacy and ethical analytics practices, future
models based on our work may serve to augment traditional mental health care procedures. More
generally, our results support the idea that computational analysis of social media can be used to
identify major changes in individual psychology.
Chapter V. Concluding thoughts

The findings presented in Chapters III and IV support the notion that meaningful psychological information can be extracted, using computational means, from social media data. Furthermore, given a suitable set of training data in which participants disclose details of their personal health histories, statistical models can be trained to predict the emergence and course of specific health conditions, based on social media data alone. Computational analyses of the coloring and social components of photographs posted to Instagram successfully predicted subsequent depression diagnoses. Language and semantic analyses of the tweets of individuals suffering from PTSD or depression revealed not only predictive indicators of these conditions, but were used to map the entire timeline of each condition, from onset to apparent recovery. The intersection of psychological and computational methods that were used to produce these findings brings both bright potential for new discoveries, and serious concerns regarding ethics and privacy. I briefly address these implications now, before concluding with a broader evaluation of the limitations, indications, and import of my work.

Two features common to all of the findings I have presented in this report is that they are a) predictive and b) specific. By predictive I mean to invoke not only the statistical definition of significance of independent variables, but also the colloquial sense of the term. The signals hidden in social media data are expressive of some mental illnesses, well before the average time to diagnosis after onset. Using the methods outlined in this report, those signals can be captured and used to infer others’ mental health states - possibly even before those individuals are themselves aware of their conditions.

By specific, I mean that the machine learning approaches I have used are capable of determining the mental health status of a specific individual person - they do not simply report
on which variables are significant in a given statistical model. To be clear, this capacity is not special to the particular algorithms I chose to use, but rather reflects a methodological approach that differs from conventional social science practices, more commonly deployed in machine learning and artificial intelligence research settings. The key observation here is that the models I have built can be used for extremely targeted analysis, at an individual level, of personal health histories - or, for that matter, of any other aspect of an individual’s personal life that can be investigated using a similar methodological setup.

The combination of these two features, prediction and specificity, opens up an entire domain of investigative research that uses big data to learn about small details. While this methodological approach has an enormous potential to help people, it also carries with it an onus to ensure that data privacy is upheld, that individual rights are respected, and that ethical research practices are not only observed, but strengthened and expanded to cover the many ramifications of this emerging computational science. As I have observed in previous chapters, there already exists a skepticism towards researchers looking to acquire and analyze social media data, caused in part by a number of egregious revelations of exactly the sort of privacy violations that people worry about (Fiske & Hauser, 2014; Kramer, Guillory, & Hancock, 2014; Lumb, 2016; Rudder, 2014). Even well-intentioned researchers often convince themselves that the public-domain status of many social media data (most notably the case with tweets) gives them license to analyze such data in any way they see fit. It is easy to overlook that the public status of many social media data does not signify users’ willingness to share private details which may have been unintentionally shared, and which may be gleaned through non-obvious technological means (e.g. via the use of computational feature extraction and machine learning algorithms). Especially when the results of such analyses, such as those reported in Chapters III and IV, can
potentially used to inform medical and clinical assessments, an extra level of caution must be exercised on the behalf of researchers when determining, for example, whether to make data sets or digital identifiers public. Gaining and upholding the trust of individuals willing to share their personal social media data is not only an ethical imperative, but will also likely determine the extent to which social computational research can be conducted at all. As more studies emerge that address similar questions and use similar methodologies to those outlined in this report, it will become critical to develop commensurate guidelines for the ethical recruitment, data collection, analysis, and publication of this novel genus of human subjects data.

**Limitations**

*Recruitment challenges and replicability issues*

When publishing new findings, it is ideal to be able to share the data used for analysis. In the case of my research, it isn’t possible to make the raw data collected public, as participants were guaranteed that their personal data, including health histories and social media posts, would not be shared. While it is possible to anonymize social media usernames, other identifiers may exist in the content of posts, such as selfies posted to Instagram, or real names included in tweets. Removing or anonymizing these potential identifiers would be extremely difficult (e.g. finding and masking names in tweets) or impossible (e.g. blurring faces in photos) to do without altering the details of the posts which are used in feature extraction and analysis. As such, the raw data cannot be made public. It might be possible to construct an aggregated dataset, free of identifiers, but that would not permit other researchers to verify the extraction and aggregation methods I used, begging the question of independent verification.

Anyone wishing to perform similar research will be required to collect new data. Data collection, especially when it involves active participant recruitment, is arguably the most
challenging aspect of social media research. There exist few viable avenues for recruitment, and many people are hesitant to share their personal data, as discussed at length above and in Chapters II-IV. In addition, platform-specific APIs are becoming more restrictive in the ways they can be used to collect information, especially as misuses and violations of API developer terms of use appear. Still, there are clearly many social media users who are willing to share their data for academic research, and future projects might do well to consider establishing a repository of data, provided by participants willing to allow their data to be used for general academic research. (Another possibility would be to establish a mailing list, populated by individuals willing to be contacted by social media researchers, which would be used to send opt-in invitations for participation in studies.) As it stands, given the technological requirements and various privacy issues involved, data collection is a considerable barrier to future research.

Sample bias and external validity

Sample bias is an important issue which limits the external validity of the findings presented here. As discussed in the closing paragraphs of Chapters III and IV, the inferential scope of my work is currently limited to a population which is characterized by the following combination of criteria:

1) Diagnosed with a mental health condition,
2) Used Twitter/Instagram at the time of diagnosis,
3) Willing to disclose health status,
4) Willing to share social media history,
5) Works on MTurk.

Ideally, the conclusions of my research would hold for all individuals who satisfy conditions 1) and 2). With the exception of 5), however, the first four characteristics are hard
requirements of all participants - 1) and 2) are prescribed by the analytical premise itself, and 3) and 4) are ethical necessities. A potential sample bias exists, as individuals willing to share personal health information and social media history with researchers may exhibit markedly different characteristics - both in the way they use social media data, and in the symptomatology of the target health condition - than individuals who are less forthcoming with personal details. One way to resolve concerns regarding sample bias would be to conduct interviews with individuals who would and would not be willing to share their personal health data and social media history with researchers, and determine whether the two groups are similar on key metrics. An assessment of the degree of similarity between both populations would help to establish whether the current analyses might reasonably be extended to a larger population.

These sources of possible sample bias are not particular to social media research. To some degree, any study which obtains informed consent from participants is open to similar criticisms. Similarly, while the use of MTurk as the sole recruitment platform raises the possibility that MTurk workers share specific characteristics that make them poor representatives of the larger population, this concern, too, is not specific to the research presented here. Future projects may find it most fruitful to conduct supplementary research to demonstrate that such biases are inconsequential, rather than attempt to design an unbiased data collection procedure. *Units of observation*

I have touched upon the issues surrounding observational units in previous chapters, but it bears revisiting as it is a methodological challenge of central importance. In building a statistical model to infer social media users’ psychological states, the biggest challenge lies in choosing appropriate units of observation. I am interested, for example, in identifying markers of depression in Instagram users. But how far through time should I tell my algorithm to cast its
assessing gaze? Photograph-by-photograph predictions invite high variance and statistical noise that may drown out a true predictive signal. On the other hand, to consider the entirety of a Instagram user’s posting history as a single, aggregated unit is just as inappropriate. An Instagram user who has recently become depressed, but whose posting history stretches back for years, would likely be overlooked by an algorithm searching for average indicators of depression. Even in the case that all posts collected for a given individual were created after being diagnosed with depression, some days will be better than others for that individual, and some posted photos may carry stronger signals of depression than others.

In between the extremes of single posts and entire histories, it would seem that some level of temporal aggregation is appropriate to capture meaningful psychological markers. I followed the precedent of De Choudhury et al. (2013) by choosing a day’s worth of posts, per each user, as the analytical unit of observation. In the Twitter study, I also conducted additional analyses with user-weeks as the units of observation, due to concerns about the instability of sentiment analysis for daily units. These choices were useful, as they allowed for more direct comparison to previous research and provided sufficient data per unit to conduct stable analysis. Nevertheless, these units are still somewhat unsatisfactory, as daily and weekly observational units do not necessarily reflect the window of time taken into account by human healthcare providers when making assessments and diagnoses.

For example, two commonly-used depression questionnaires ask respondents to only refer to their moods over the past two weeks (see Radloff, 1977, and Spitzer, Kroenke, & Williams, 1999). Would it be more appropriate, then, to adjust future analyses to consider two-week units of observation, rather than daily or weekly units? Or perhaps it would be most appropriate, for the purposes of direct comparison to conventional mental health screening
procedures, to only consider the two weeks of social media posts made prior to diagnosis when fitting a predictive model? The former solution would require adjusting observational units per each condition, based on the period of time considered when making diagnoses for a given condition - assuming such a standardized time period exists. The latter solution would make it impossible to generate illness timelines via state-space analyses like the Hidden Markov Model approach used in the Twitter study, meaning that separate observational units would need to be created for classification and time series models.

An additional concern regarding user-days as units of observation is that they exhibit structural dependency. Many methods of statistical inference, particularly common parametric algorithms such as linear and logistic regression, operate on the assumption of independence of observations. When observational units are nested by users, as is the case with the user-day units used in my analyses, they may exhibit structural covariance within their hierarchical grouping (i.e. User A’s posts will have more in common with each other than with the posts of User B), violating common statistical assumptions of independence. In cases where models do carry such assumptions, this dependency can be accounted for by adopting a hierarchical framework, where observations are explicitly allowed to share variance per their grouping units (here, individual users). Other models, such as Random Forests, do not make independence assumptions, and so are unaffected by observational groupings. In either case, the choice of grouped units of observation require extra care in selecting appropriate statistical models for analysis.

Thinking about these questions highlights the dangerous potential for gross categorization which is inherent even in conventional psychological diagnostic practices. One rarely hears people say “My last two weeks were clinically depressed”, but rather “I am clinically depressed”, even though the former may be a truer representation of the information taken into account by a
diagnosing physician. As such, the consideration of observational units when screening for
mental illness may not only be a statistical concern, but one that affects traditional screening
procedures as well. Future research might benefit from closer consultation with mental health
care providers in selecting the most effective units of observation, to allow for close comparison
against typical human diagnostic accuracy.

**Future directions**

The research described in this report joins a nascent literature in the field of
computational social science, with a focus on predictive health screening. As the field continues
to mature, there are several areas in which it will need to improve or innovate in order to advance
both discovery and adoption. Here, I briefly consider three such areas for improvement.

*Appropriate benchmarking against human performance*

Contextualizing the value of a computational result in an appropriate and meaningful way
can be a challenging task. When a computational alternative to a traditional human task is
designed, for example, there comes with it a natural inclination to compare machine and human
performance. In chapters III and IV, for example, I provided data on human physicians’
diagnostic accuracy when assessing depression and PTSD. As noted in those chapters, however,
it isn’t appropriate to make a direct comparison of accuracy metrics between human physicians
and my predictive models, for a number of reasons. First, results are based on two different
samples. Second, the units of observation are different. Third, the task being evaluated isn’t the
same, as human physicians are faced with evaluating their patients for any number of health
conditions, where my computational models only know they must assign a zero or one (i.e.
indicate the presence or absence of a single condition). The two tasks are different enough so
that even an informal comparison may yield objections from an informed audience.
Given all these disparities, it might be proposed that it is better to evaluate computational results with strictly formulaic benchmarks. In the case of null hypothesis significance testing, comparison against a null model would offer a relative assessment of model strength. In the context of machine learning classifiers with class imbalances, comparing accuracy metrics against a naive model which predicted the majority class for all observations would offer a similar point of contrast. The value of such comparisons is questionable, however. A null, or naive, model may represent a logical counterpoint to a predictive model, but may also be eminently impractical. For example, under what circumstances would a physician ever assume that all people are healthy, simply because most people are healthy? This is analogous to the naive model of an imbalanced-class classifier. Just as it wouldn’t make sense to compare a new diagnostic method against an unrealistic precedent, so it also makes little sense to proclaim a classifier better than a naive model which is unreflective of reality.

Despite the difficulties inherent in drawing comparisons between typical human diagnostic performance and computational assessment, I advocate for framing analytical results in a practical context. When reporting such comparisons, care should be taken not to draw inappropriate conclusions, but the basic premise of benchmarking against human performance seems both reasonable and appropriate to me. An ideal solution for future research would be to have human physicians and computational models make assessments on the same participant sample (although this might introduce considerable challenges with respect to recruitment and data collection). 

*Augmenting bulk ingestion methods with pre-built models*

Data collection involving active recruitment procedures is markedly slower and generally results in smaller samples than the bulk ingestion approach described in Chapter II. The main
reason to use an active recruitment design is to be able to gather specific information from individual participants which may not be available from querying their social media data. While most studies involving active recruitment will likely be bound from sharing the raw data collected by data privacy agreements, the fitted predictive models which result from those studies will almost certainly be shareable, as a fitted model is often no more than a vector of numerical weights. Applying pre-built models to data collected via bulk ingestion may be a powerful means of validating and extending models with sizable datasets. This approach might also highlight potential limiting factors of models fit from active recruitment samples, which might then be used to refine modeling strategies and hypotheses in future iterations of active recruitment. This synthesis between the precision of active recruitment and the sheer quantity of data available in bulk ingestion may serve as an important feedback cycle as research in this field continues to advance.

*Introducing experimental design into computational social science*

A distinguishing feature of computational social science research to date is that most studies are observational. The value of such research has so far been found more in mining insights from existing data, rather than creating new data (and my own results are no exception to this trend). It seems to me that this bias towards observational studies is not a result of some insurmountable technological obstacle, but rather because many computational social scientists are simply not trained in experimental research (and, conversely, not many experimentalists are trained in computational science). In fact, purveyors of social media, including Facebook and OKCupid, have publicly acknowledged their research in experimentally manipulating members’ data to refine their respective products (Kramer, Gu Ellory, & Hancock, 2014; Rudder, 2014).
Here I briefly provide a few suggestions and recommendations for those considering an experimental approach.

1. Successful experimental designs on social media will likely be longitudinal.

   Analysis of social media data often assumes that data are available in sufficient quantity for stable analysis. In the case of text sentiment analysis (e.g. evaluating the happiness of Twitter posts), unigram sentiment instruments are often unstable when applied to a set of text with only a few words. As such, any experimental manipulation which relied on analysis of subjects’ subsequent social media activity might need to wait for subjects to post a sufficient quantity of material before analysis can take place. This consideration may limit the type of treatment or experimental manipulation that can reasonably be expected to produce an effect in this setting (e.g. a temporary priming effect might wear off before enough data are produced to be analyzable).

2. Study participants should meet an average posting threshold.

   Related to the first point above, experimental subjects should be selected, in part, for their active statuses on the social media platform of interest. People who post only irregularly may weaken or distort the signal produced by an experimental manipulation. This selection procedure introduces its own bias, however, as frequent posters may be qualitatively different than infrequent posters, and should be considered when designing this kind of research.

3. Downstream exposure effects should be accounted for with network analysis.

   Depending on the experimental manipulation, its effects may spread to other social media users, in addition to experimental subjects, by virtue of subjects’ communication with others in their social networks. This potential for down-network transmission may be seen as a “bug” or a “feature” of the design, so to speak, and should be explicitly accounted for in an experimental
design. For example, in recruiting a control group of social media users who have not been exposed to an experimental manipulation, it may be advisable to exclude individuals within a few degrees of connection to subjects in the experimental group on the social media platform being studied. Similarly, the same possibility of down-network transmission may itself become the subject of an experimental hypothesis, as has been demonstrated in a number of graph-analytic approaches to social network analysis (see, for example, Christakis & Fowler, 2010).

**Motivating future research in psychology**

Computational social science, as a field, counts very few psychologists among its leading researchers (Lazer et al., 2009). As far as I can tell, there are two main reasons for this: 1) computational science tends to use more techniques from applied math, statistics, and computer science than most psychologists are familiar with, and 2) the problems addressed by computational science are perceived as either applied or methodological research, and so are seen as uninteresting to researchers mainly concerned with advancing psychological theory. The research presented in this report does little to dispel these notions. It makes novel methodological contributions to the fields of psychological and computational science. It also suggests promising future applications for mental health screening in health care. And it accomplishes these goals with a suite of tools culled from computer science and machine learning. While I personally envision a more syncretic psychological science on the near horizon, in which psychologists will receive training in a much wider variety of methods and disciplines than is currently common, I also recognize that the computational social science formulation will likely turn off many psychological scientists from continuing this line of research, due to the concerns about novelty and interestingness noted above. Granted, this will certainly not hold true for all - but for those for whom the prospect of investing time conducting
“methods” or “applied” research seems unappealing, I will close by offering a few general observations in an attempt to persuade otherwise.

When speaking with other psychological scientists, I frequently encounter an aversion to applied research. Applied research, so the arguments go, addresses less interesting questions than does basic research, and is positioned further downstream in the thought chain of psychological science. Medin (2012) challenged these assumptions - noting, in particular, that applied work has frequently served as inspiration for theoretical discoveries. To add to that point, I argue that the research I have presented here is actually an example of basic research which happens to be well-aligned with practical application - a twofer, so to speak. Making a novel contribution to one’s field, and presenting a blueprint for applied work, need not be mutually exclusive tasks. The combined themes involved in these studies - social media, mental health, and predictive screening technologies - are both new and useful enough to satisfy both basic and applied researchers.

The focus on methods in this research may also be off-putting to some researchers who are more interested in the why than the how of psychological science. I imagine this is both because many of the methods, including programmatic data collection, Bayesian statistics, machine learning, and state-space models, may be unfamiliar. In general, methodological advances in psychological science are rewarded less (or seen as less interesting) than theoretical advances. With regards to the first point, unfamiliar methods, that is a problem easily solved by finding computational scientists interested in collaboration! (More advisable still would be to gain personal competency with more computational methods, but this admittedly requires a larger up-front investment.) The latter point, regarding theory versus methods, is, to an extent, born of an erroneous assumption of independence between these two categories. Consider the
old adage, “When all you have is a hammer, everything looks like a nail.” It is not unreasonable to presume that the thought process undertaken when searching for new hypotheses is constrained, to a degree, by the available means of testing them. As such, expanding the knowledge base of available methods may actually expand the scope of theoretical advances, by virtue of allowing more researchers to think about more ways to test out ideas.

In addition to the conceptual argument, that expanded methods can lead to expanded theory, it is also notable that some interesting theoretical questions can arise as a by-product of methodological research. For example, in the report on Instagram and depression (Chapter III), one finding was that depressed Instagram users tended to post photographs with fewer human faces. I made a speculative connection to the work of Rude et al. (2004), who found an increased use of self-focused language in depressed individuals (which in turn was based on Pyszczynski and Greenberg’s (1987) theory of depression and self-focused attention. To my knowledge, no corresponding hypothesis regarding self-focused imagery and depression has been examined. Either of the outcomes of such an experiment - that a link between self-focused imagery and depression does or does not exist - could constitute useful theoretical contributions to contemporary models of depressive behavior.

Another interesting theoretical question refers to the time series graphs of the onset and course of depression and PTSD (Chapter IV): Why are these illnesses so clearly observable in the public online communications of affected individuals? Are these indicators as clearly present in other, more private modes of communication? Further investigation into the ways people disclose indicators of mental illness (whether purposeful or not) in public versus private settings also may be of theoretical value.
In summary, I strongly advocate for psychological researchers to include more methods from computational social science, with the aim of improving both methodological and theoretical contributions to psychological science. Given the time constraints imposed on many graduate research programs (and even for untenured professors), it may seem a tenuous proposition to take time out from generating research to learn more methods, but I’d argue it is worth it. As a field, computational social science would greatly benefit from the expertise of psychological scientists, and accordingly there are many opportunities for motivated researchers to make a considerable impact in both computational and psychological science.
References


http://doi.org/10.1371/journal.pone.0026752

http://doi.org/10.1001/archpsyc.57.3.217


http://doi.org/10.1073/pnas.141462611


http://doi.org/10.1177/0959354307086923


http://doi.org/10.1109/MTS.2012.2225462


http://doi.org/10.1001/archpsyc.62.6.617


http://doi.org/10.1126/science.1248506


Appendix

I. Hacking Qualtrics (Chapter II)

The Qualtrics survey software platform served as the main vehicle for collecting information from participants besides their social media data. Screening questions, mental health scales, and questions about mental health history, were all presented to participants using Qualtrics surveys. Once participants came to the point in the survey where they are asked to provide their Twitter/Instagram username, and to agree to share their social media history, there needed to be some way of connecting the live survey window (a web page) with both their social media account (via a separate web page) and with the backend data processing infrastructure I built on my own private server (necessary to keep track of the various participants, social media usernames, and data that were involved across all studies). All of this information passing needed to happen in real-time, without disconnecting participants from the survey. See Figure 13 for an example of the three separate web pages a survey participant is brought to, in order to connect their survey responses with their social media data (and for my background programs to validate and store this information).
While Qualtrics is an extremely capable piece of software, it is designed to operate mainly within its own ecosystem, and currently does not support this kind of cross-platform, cross-server connectivity. Fortunately, there are two features of the Qualtrics software that allow for customization and data processing that goes beyond the standard user interface: (1) JavaScript injection into survey documents, and (2) the Qualtrics API. JavaScript is a programming language commonly used in web applications, and enables a protocol, called AJAX, that permits real-time communication between different servers. Injecting custom JavaScript into study surveys allowed me to link Qualtrics with Twitter and Instagram without forcing participants to leave the survey page. This in turn made it possible for participants to log
into their social media profiles, grant access for my application to collect their data, and return to Qualtrics to complete participation\textsuperscript{34}. Qualtrics’ API allowed me to write a program that interfaced directly with Qualtrics’ data servers, which automated the time-consuming process of retrieving participant data from Qualtrics, converting it to a file, and then copying file contents into a database. Without these customization capabilities, linking survey responses with social media data (or any kind of Qualtrics-external server activity) would have been difficult, if not impossible, to carry out.

\textsuperscript{34} For the more technically-minded reader, wondering about Cross-Origin Resource Sharing issues: CORS obstacles were resolved by acquiring a Comodo SSL certificate for my backend server, which handled request passing between Qualtrics, Twitter/Instagram, and my data collection application.
II. MCMC Diagnostics (Chapter III)

Figure 14. Trace and density plots for All-data model MCMC simulations.
Figure 15. Autocorrelation plot for All-data model MCMC simulations. First chain only is displayed for conciseness (second chain output is nearly identical).
Figure 16. Trace and density plots for Pre-diagnosis model MCMC simulations.
Figure 17. Autocorrelation plot for Pre-diagnosis model MCMC simulations. First chain only is displayed for conciseness (second chain output is nearly identical).
Figure 18. Trace and density plots for Ratings model MCMC simulations.
Figure 19. Autocorrelation plot for Pre-diagnosis model MCMC simulations.