Predicting Mental Health from Followed Accounts on Twitter

Cory K. Costello, Sanjay Srivastava, Reza Rejaie, & Maureen Zalewski

University of Oregon

Author note:

Cory K. Costello, Sanjay Srivastava, and Maureen Zalewski, University of Oregon, Department of Psychology, 1227 University of Oregon, Eugene, OR, 97403. Reza Rejaie, University of Oregon, Department of Computer and Information Science, 1202 University of Oregon, Eugene, OR, 97403.

This material is based on work supported by the National Institute of Mental Health under Grant # 1 R21 MH106879-01.

Correspondence should be sent to: Cory K. Costello, Department of Psychology, 1227 University of Oregon, Eugene, OR 97403, email: ccostell@uoregon.edu.

Abstract

The past decade has seen rapid growth in research linking stable psychological characteristics (i.e., traits) to digital records of online behavior in Online Social Networks (OSNs) like Facebook and Twitter, which has implications for basic and applied behavioral sciences. Findings indicate that a broad range of psychological characteristics can be predicted from various *behavioral residue* online, including language used in posts on Facebook (Park et al., 2015) and Twitter (Reece et al., 2017), and which pages a person 'likes' on Facebook (e.g., Kosinski, Stillwell, & Graepel, 2013). The present proposal seeks to examine the extent to which the accounts a user follows on Twitter – their Twitter friends – can predict individual differences in self-reported anxiety, depression, post-traumatic stress, and anger. Studying Twitter friends offers distinct theoretical and practical advantages for researchers, including the potential for less overt impression management and better capturing passive users. By incorporating best practices in open science and machine learning, we aim to provide unbiased estimates of predictive accuracy for predicting Mental Health from Twitter friends. Our findings will have implications for theories linking psychological traits to behavior online, applications seeking to infer psychological characteristics from records of online behavior, and for informing discussions of how such applications could affect users' privacy.

Predicting Mental Health from Followed Accounts on Twitter

Stable psychological characteristics are expressed behaviorally in many domains, including online, where they often leave more or less permanent digital records in their wake. The extent to which stable individual differences in mental health are expressed online, imprinted in corresponding digital records, and ultimately recoverable from these records has wide-ranging implications for basic and applied behavioral sciences. Inferring individuals' mental health status from online records with an appreciable degree of accuracy could accelerate advancements in clinical science, easing the burdens for researchers and participants imposed by traditional survey-based research. In time, such approaches could be developed into tools useful for clinical practice and public health. At the same time, the promise of inferring mental health from digital records of behavior is accompanied by potential threats to individuals' privacy, as such tools could be used to infer a person's mental health without their explicit consent. Given both the promise and risks, we need to better understand how mental health is reflected in, and recoverable from, digital records of online behavior. Our focus here is on inferring depression, anxiety, anger, and post-traumatic stress from the accounts users choose to follow on the popular online social network (OSN), Twitter.

Psychological Traits can be Inferred from Digital Records

The theory of behavioral residue holds that one by-product of the expression of traits is the accumulation of lasting residual traces of past behavior in the physical or digital spaces a person occupies. Early work demonstrated that human judges could infer psychological traits from behavioral residue in physical living and working spaces with considerable accuracy (Gosling, Ko, Mannarelli, & Morris, 2002). More recently, researchers have trained machine learning algorithms to do so with behavioral residue found in OSNs such as Facebook (Kosinski,

Stillwell, & Graepel, 2013; Park et al., 2015; Schwartz et al., 2013). Behavioral residue in OSNs has included linguistic content (e.g., Facebook status updates; Park et al., 2015; Schwartz et al., 2013) and which pages a person has 'liked' on Facebook (Facebook-like ties; e.g., Kosinski, Stillwell, & Graepel, 2013), both having demonstrated considerable predictive accuracy. Indeed, the accuracy of inferences based on Facebook-like ties can even exceed that of knowledgeable human judges (Youyou, Kosinski, & Stillwell, 2015). We focus here on using behavioral residue from Twitter, an OSN that differs from Facebook in ways relevant to basic psychological theory, public health applications, and privacy concerns.

Twitter is an OSN service and microblogging platform used by approximately 24% of US adults (as of January 2018; Pew Research Center, 2018). Users post short messages of no more than 280[1] characters called "tweets" that other users can see, respond to, share (called "retweeting"), or react to (via a "like" button). Unlike Facebook, accounts are public by default, and most users choose to keep their accounts public; Twitter does not release the percentage of public accounts, but a 2009 report found that 90% of accounts were public, with a trend towards even fewer private accounts (Moore, 2009). The public nature of Twitter makes it an especially interesting setting for the present investigation for two reasons. First, its public nature eases the burden of collecting users' data: one of several off-the-shelf Python (e.g., *Tweepy*; Roesslein, 2009) or R libraries (e.g., *twitter*; Gentry, 2015) can be used to download any of these many public accounts' data, including their recent tweets, whom they follow, and who follows them. Thus, there is at least one fewer barrier to people outside of the Twitter company for implementing beneficial (e.g., public-health) or harmful (e.g., discriminatory) applications on Twitter than less public-facing OSNs like Facebook. Second, its public nature could affect the

[1] Prior to November 2017, tweets were limited to 140 characters.

relative candor of behavior on Twitter, since efforts to manage others' impressions can be stronger in more public settings (Leary & Kowalski, 1990; Paulhus & Trapnell, 2008).

Previous work has attempted to infer or predict psychological traits from behavioral residue on Twitter, focusing primarily on linguistic analyses of tweets. This growing body of work demonstrates that tweets can be used to predict a wide range of psychological characteristics, including self-reported personality traits, affective states, depression, post-traumatic stress, and the onset of suicidal ideation (Coppersmith, Harman, & Dredze, 2014; De Choudhury, Counts, & Horvitz, 2013; De Choudhury, Gamon, Counts, & Horvitz, 2013; De Choudhury, Kiciman, Dredze, Coppersmith, & Kumar, 2016; Dodds, Harris, Kloumann, Bliss, & Danforth, 2011; Nadeem, Horn, & Coppersmith, 2016; Park, Cha, & Cha, 2012; Qiu, Lin, Ramsay, & Yang, 2012; Reece et al., 2017). Although the heterogeneity in how mental health is measured can make interpretation challenging, in broad terms these previous studies suggest that behavior on Twitter relates meaningfully to psychological traits. In contrast to the emphasis on linguistic analyses, there has been relatively little work using network ties on Twitter to predict psychological traits. The few attempts have looked at abstract structural characteristics of ties (e.g., tie counts or social network density; Golbeck, Robles, Edmonson, & Turner, 2011; Quercia, Kosinski, Stillwell, & Crowcroft, 2011) rather than treating the specific accounts to which a user is tied as meaningful. We focus here on the specific accounts that users follow.

Ties or connections on Twitter are directed, meaning that users can initiate outgoing ties (called "following" on Twitter) and receive incoming ties (called "being followed" on Twitter) which are not necessarily reciprocal. In keeping with the terminology of Twitter's Application Programming Interface (API), we refer to the group of users that a person follows as their *friends*

and the group of users that follow a person as their *followers*. While both ties are likely rich in psychological meaning, we focus on friends in the present investigation for several reasons.

First, a user has nearly complete control over the accounts they follow, making friends a more direct product of the user's own behavior. While most of a users' followers likely reflect their own behavior and relationships, some might be unrelated (e.g., spam accounts, bots, users looking for reciprocated following, etc.), increasing the relative noise among followers (vs. friends). Second, following accounts is the primary way users' curate their feed – what they see when they log into the app – and so their choice of friends likely reflects the information they are seeking out on Twitter. In this way, following accounts on Twitter is similar to liking pages on Facebook, a behavior which has been previously demonstrated to robustly predict psychological characteristics (Kosinski et al., 2013; Youyou et al., 2015). Third, an important practical consideration for a predictive modeling approach is that friends on Twitter often include famous brands, celebrities, politicians, or other *high in-degree accounts*, which appeal to similar interests or motivations in many users. Users are thus likely to share some friends even in moderately small samples, whereas they may have no followers in common because there aren't parallel *high out-degree accounts* that appear in many users' follower networks. Consequently, friends are far less likely to be zero-variance predictors than followers in moderately-sized, random samples.

Predicting Mental Health from Twitter-Friend Ties

Twitter-friend ties are an important next step in studying behavioral residue online for both theoretical and practical reasons. In contrast to tweets, Twitter-friend ties are *not* explicit signs or displays intended to be consumed by an audience of other people, and so they may be less subject to impression management goals. For this reason, Twitter-friend ties may be

especially apt for predicting more evaluative psychological traits like mental health status. Likewise, people may be unaware of how much they are divulging with their selection of Twitter friends, heightening the relevant privacy concerns. For example, imagine a Twitter user who wants to present as less depressed than they truly are. They may be well aware that they should avoid writing tweets that express negative emotionality (typically the best linguistic cues for depression; see e.g., Reece et al., 2017), but they may not think to tailor their selection of friends to serve this impression management goal.

Another important advantage of Twitter-friend-based assessments is that they should work well with both *active* and *passive* users, which are distinguished in terms of the extent to which they actively engage with others (e.g., tweeting, commenting, etc.) or passively consume content (e.g., read tweets posted by their friends; Verduyn et al., 2015). Passive users tweet less often by definition, so tweet-based predictions are less suited for them. But if they are using Twitter to passively consume information, they will still follow other accounts and thus establish a set of Twitter-friend ties. Incorporating passive users may be especially advantageous for improving predictive accuracy when examining mental health such as depression and anxiety, because symptoms such as withdrawn behavior, indecision, or worry, may manifest as more passive than active twitter behavior; users who would generate insufficient data for an analysis of posted tweets may still follow a sufficient number of accounts to analyze friends.

Although the psychological meaning of Twitter friends is perhaps less immediately obvious than the psychological meaning of tweets, there are several reasons to suspect that it may be rich, and that people are therefore unknowingly disclosing sensitive information about psychological traits like anxiety, anger, depression, and post-traumatic stress through their friend networks. Individuals' mental health could affect which accounts they choose to follow in

several ways. One theory anticipating this is *homophily*, which holds that people like and therefore seek out others who are similar to themselves. For example, relatively depressed individuals would be anticipated to differentially follow other similarly depressed individuals or accounts. Informative to the present work, homophily has been consistently observed (offline) for individual differences in emotion (Anderson et al., 2003; Watson et al., 2000a, 2000b, 2004, 2014) and mental health status such as depression (Schaefer, Kornienko, & Fox, 2011), and has recently been found in OSN friendship-ties (Youyou, Stillwell, Schwartz, & Kosinski, 2017). Mental health could also affect Twitter-friend ties if selecting Twitter friends reflects strategies to regulate one's emotions via situation selection (Gross, 2002). For example, a person who is relatively more depressed may seek out especially positive content on Twitter to upregulate positive emotions.

The reverse causal direction is also possible and Twitter friends could affect individuals' mental health, either via emotion contagion processes – which have been observed in online social networks (Kramer, Guillory, & Hancock, 2014) – or through other mechanisms. Both could affect each other complementarily, creating a person-environment transaction whereby people select themselves into networks that reinforce their existing mental health (Buss, 1984). For example, negative world views are a psychological component of depression, which may be expressed on twitter by seeking out friends that reaffirm this negative world view, which may in turn exacerbate depression symptoms. We will not be able to distinguish between these different possibilities in our data, but each would facilitate friend-based predictive accuracy.

The Present Study

The present study will examine the extent to which psychological traits relevant to a person's mental health and well-being can be inferred from their Twitter friends. We will focus

on self-reported depression, anxiety, anger, and post-traumatic stress, providing a relatively broad range of important mental health constructs. We incorporate best practices in psychometrics, open science, and machine learning. Our outcomes are measured with well-validated, psychometrically sound measures, which should enhance the predictive accuracy and explanatory utility of the results. To ensure unbiased estimates, we will incorporate pre-registration and a holdout sample into our data analysis workflow, where we will first perform all model training and selection in part of the data, pre-register our final models (in a publicly-available, timestamped registration), and finally test them on the holdout sample. In doing so, our results will provide a highly rigorous test of the extent to which a broad range of mental health constructs are reflected in Twitter friend ties.

Method

Participants and Procedure

Data collection was approved by the University of Oregon Institutional Review Board (Protocol # 12082014.013) and was conducted in a manner consistent with the ethical treatment of human subjects. We collected data from the Spring of 2016 until the Fall of 2017, recruiting participants primarily from the "r/beermoney" and "r/mturk" Reddit communities, with additional participants from the University of Oregon Human Subjects Pool (UOHSP), Amazon's Mechanical Turk (mTurk), and Twitter advertising (using promoted tweets).

Our inclusion criteria required participants to provide an existing unlocked Twitter account, currently reside in the US, and to meet minimum thresholds for being an active Twitter user. Active twitter users were defined as having a minimum of 25 tweets, 25 friends, and 25 followers. Using two-stage prescreening, we attempted to first screen participants for eligibility before they completed the main survey; participants had to affirm that they met the inclusion

criteria before they proceeded with the main survey. However, since participants could erroneously state that they met the inclusion criteria, each participant was individually screened by the first author to verify that they indeed met the criteria, and to further assess whether the Twitter handle belonged to the participant whom provided it. This consisted of manually searching each Twitter account provided, ensuring it met the activity thresholds, and assessing whether the account provided was obviously fake (e.g., one participant provided Lady Gaga's account and was subsequently excluded). When it was especially difficult to verify that the accounts provided belonged to participants, we contacted them to confirm that they indeed owned the account they provided by direct messaging our lab's Twitter account from the account they provided.

The total number of participants recruited through each mechanism as well as the subset that were verified as meeting the inclusion criteria are shown in Table 1. As shown in Table 1, this process led to a total of $N_{eligible} = 762$ accounts that we were able to verify met our inclusion criteria. Ineligible prescreen participants contained a mixture of participants who did not provide an existing Twitter account, participants who provided an account that they did not own (e.g., Lady Gaga's account), participants whose Twitter account did not meet the activity thresholds, and participants that provided an eligible but locked account.

In all recruitment methods, participants were able to click a link that took them to the Qualtrics survey where they provided their Twitter handles, answered some questions about their Twitter use, completed several self-report measures (described below), and finally completed basic demographics questions. At the end of the survey, participants were thanked, and compensated either with an Amazon gift card or physical check for $10 or with course credit for participants recruited through the human subjects pool.

We then downloaded each eligible participant's full friends lists from Twitter's API. Of the 762 eligible accounts, we were unable to get friends lists from 101 participants, one user was suspended, and the remaining 101 either deleted, locked, or changed the handle[2] of their accounts in the intervening time between screening for eligibility and collecting friends lists from Twitter's API. This resulted in the final sample for the present study of $N = 661$ active Twitter users.

Measures

Participants completed a series of psychometric scales, including the measures of anxiety, anger, depression, and post-traumatic stress relevant to the present study. Anxiety, anger, and depression were measured using short-forms of the Patient-Reported Outcomes Measurement Information System (PROMIS; Pilkonis et al., 2011) questionnaires. These measures asked participants to indicate the frequency of symptom occurrence on a scale ranging from 1 (never) to 5 (always). The anxiety scale consisted of eight short statements (e.g., 'I felt fearful'), the anger scale consisted of five short statements (e.g., 'I felt like I was ready to explode'), and the depression scale consisted of eight short statements (e.g., 'I felt worthless'). Post-traumatic stress was measured using the well-validated 10-item Trauma Screening Questionnaire (TSQ; Brewin et al., 2002), which asks participants to indicate whether or not they have experienced 10 symptoms ('upsetting dreams about an event') at least twice in the past two weeks.

Planned Analyses

To date, we have kept ourselves blind to the data. We have not conducted any analyses in any of the collected data, only manually inspecting it to ensure that it was collected correctly.

[2] Our workflow consisted of looking up users in Twitter's API based on the handle they provided using application-only authentication.

This approach enabled us to write an unbiased pre-registration before running any analyses (Srivastava, 2018). Unless otherwise noted, all analyses will be conducted in R (version 3.6.1 or later; R Core team, 2018). Figure 1 depicts our planned general data analysis workflow (1a), model training workflow (1b), and model testing workflow (1c). In broad strokes, our aim is to train and select a predictive model for each mental health variable that 1) maximizes out-of-sample predictive accuracy, 2) guards against over-fitting, and 3) is interpretable, providing insight into how and why mental health may be recoverable from friend relations on Twitter. To meet these aims, we will partition our data into a training and holdout sample, perform all feature selection, data reduction, model training, and model selection on the training sample, and use the holdout sample only to assess accuracy of the final model. Our aims led us to choose four modelling approaches (detailed below) as our candidate models. We selected these four approaches based on a combination of what has worked previously with similar data in published studies (e.g., Kosinski et al., 2013) and our own feasibility studies (described later), techniques that are well-suited to Twitter friends data (e.g., algorithms that work well with sparse predictors), and potential interpretability. The specific rationale for each of the four approaches is detailed below.

Data partitioning. As shown in Figure 1a, we will first split the final sample (N = 661) into a training and holdout (testing) set using the Caret package in R (version 6.0-80; Kuhn et al., 2018). The training and holdout samples will consist of roughly two-thirds ($n_{training}$ = 438) and one-third ($n_{holdout}$ = 223) of the data respectively. All feature selection, data reduction, model training, estimation, and selection will be determined from the training data. The final model(s), trained and selected within the training data, will be tested on the holdout sample to get an unbiased estimate of out-of-sample accuracy.

Model training. Figure 1b shows our planned model training workflow and approach. As seen in Figure 1b, we will first conduct explicit feature selection, and then we will train and evaluate models using four different approaches (under each of the three feature selection rules). Each mental health variable will be modelled separately, and so the model trained and selected for one construct (e.g., depression) may differ in every respect (approach, feature selection threshold, hyperparameters, parameters) than the model trained and selected for another construct (e.g., post-traumatic stress). All models will be trained, tuned, and evaluated (within-training evaluation) using k-fold cross-validation. This splits the data into *k* random subsets called *folds*, trains the data with k-1 folds, and tests the model's performance on the k^{th} fold; this is repeated until each fold has been the *test* fold. We plan to set k to 10, which is commonly recommended (Kosinski, Wang, Lakkaraju, Leskovec, 2016). This procedure is an efficient means for reducing overfitting during model training and selection (Yarkoni & Westfall, 2017).

Explicit feature selection. The data being used to predict mental health variables will be structured as a user-friend matrix, where each row is an individual user, each column is a unique friend followed by some user(s) in the sample, and cells are filled in with 1's or 0's indicating whether (1) or not (0) each unique user follows each unique friend. The number of unique friends, or *features* (also sometimes called predictors), in the data is likely to exceed what is computationally feasible or efficient. Moreover, accounts followed by few users are unlikely to be practically useful. At the extreme, uniquely followed accounts are effectively zero-variance predictors and therefore useless for most modeling and data reduction techniques. As such, the first step of our model training will consist of minimal feature selection, pruning friends from the data that have few followers in our data (see Figure 1b) analogously to Kosinski and colleagues (2013) approach to Facebook likes. The optimal threshold for feature selection in this data is not

yet known, so we will try three values, eliminating friends followed by fewer than 3, 4, and 5 of the participants in our data; the minimum of 3 was chosen because any lower often led to model convergence issues in the preliminary analyses (described below). The feature selection rule that shows the best training performance (see model selection section below) will be used in the test data. Some modelling approaches will perform further feature selection and/or data reduction; they are described alongside the corresponding approach below.

Modelling approaches. As shown in Figure 1b, we will compare four different modelling approaches: Relaxed LASSO, Random Forests, Supervised Principal Components Analysis (Supervised PCA), and two-step Principal Components Regression (PCR) with ridge regularization. Each is described in greater detail below.

Mirroring Youyou and colleagues' (2015) approach to predicting personality from Facebook likes, we will train a model predicting each mental health variable with a variant of LASSO regression on the *raw* user-friend matrix, treating each unique twitter friend as a predictor variable. Classic LASSO is a penalized regression model which, like ordinary least squares (OLS), seeks to minimize the sum of squared errors and additionally seeks to minimize a function of the sum of absolute beta (regression coefficient) values (i.e., the L1 penalty, $\lambda * \sum_{j=1}^{j} |\beta_j|$, where λ is a scaling parameter that determines the weight of the penalty). However, classic LASSO is known to perform poorly in contexts like these, with many noisy predictors (Meinhausen, 2007)[3]. Meinhausen (2007) developed relaxed LASSO to overcome this issue, by separating LASSO's variable/feature selection function from its regularization (shrinkage)

[3] We confirmed that classic LASSO is a poor fit for this kind of data; some models fit with classic LASSO during the preliminary analyses (described below) produced predicted values outside of the bounds of the observed data (indicated by RMSE values in the millions; see LASSO section of 'predictive_modelling_both_samples.html' at the projects' osf site: https://osf.io/ky7u3/). Relaxed LASSO did not suffer from such issues.

function. Essentially, it runs two LASSO regressions in sequence; the first performs variable selection, selecting *k* predictors (where *k* is ≤ total number of predictors *j*) based on scaling hyperparameter λ, and the second performs a (LASSO) regularized regression with the remaining *k* variables, shrinking the parameter estimates for the reduced variable set based on scaling hyperparameter ϕ (see Figure 2b). Relaxed LASSO, like classic LASSO, can be difficult to interpret when features are correlated, which may or may not be the case with Twitter friends in our data.

The second approach uses the Random Forests algorithm on the raw user-friend matrix (see Figure 1b). Random Forests works by iteratively taking a subset of observations (or cases) and predictors, building a regression tree (i.e., a series of predictor-based decision rules to determine the value of the outcome variable) with the subset of predictors and observations, and averaging across the iterations. It is thus an *ensemble* method, which avoids overfitting by averaging across many models trained on a subset of participants and features. It works well with *sparse* predictors (Kuhn & Johnson, 2013), making it a promising candidate for Twitter friends. Like LASSO, interpretation can be difficult in the presence of correlated features. Although Relaxed LASSO and Random Forests are promising, their difficulty with correlated features could be problematic if Twitter friends are highly correlated. Our third and fourth approaches were chosen in part because they are more robust to this potential issue.

Our third approach will be Supervised Principal Components Analysis (sPCA), which first conducts feature selection by eliminating features that are below some minimum (bivariate) correlation with the outcome variable, and then performs a Principal Components Regression (PCR) with the remaining feature variables; both the minimum correlation threshold and number of components to extract are traditionally determined via cross-validation (Bair, Hastie, Paul, &

Tibshirani, 2006; see Figure 1b). Interpretation tends to be relatively straightforward, even with correlated predictors, making it a promising candidate for the present aims.

Finally, mirroring Kosinski and colleagues (2013)[4], we will conduct a two-step PCR with ridge regularization, first conducting an unsupervised PCA on the user-friend matrix and using the resulting (orthogonal) components as predictors in a Ridge regression; we will extract the number of components that corresponds to 70% of the original variance. Ridge regression is similar to LASSO but seeks to minimize the sum of squared coefficient values (i.e., L2 penalty; $\lambda * \sum_{j=1}^{j} \beta_j^2$) instead; it also shrinks coefficients to be closer to zero, but tends to allow more (small) non-zero coefficients. Ridge, like LASSO, provides relatively interpretable solutions when predictors are uncorrelated, which is the case with orthogonal principal components.

Model selection. As mentioned above, all models will be trained using the training data, and each model's training performance will be indexed via root mean squared error (RMSE) and the multiple correlation (R) from 10-fold cross-validation. Although machine learning approaches tend to prioritize predictive accuracy over interpretability (see Yarkoni & Westfall, 2017); we aim to maximize both to the extent possible. As such, we will select our final model based on both (quantitative) model performance criteria (minimal RMSE, maximal multiple R) and (qualitative) interpretability. Note that in addition to RMSE/R for the best performing model, we will also consider the spread of training results (e.g., we may choose a model that did not have the best single performance, if it has less variability in performance). We will therefore

[4] One major difference between Kosinski and colleagues' (2013) and our approach is that we plan to use PCA instead of singular value decomposition (SVD). PCA is special case of SVD where the data matrix is first centered; because of this centering, it tends to be less computationally efficient but more interpretable. Our feasibility analyses (described below) indicate that PCAs are indeed feasible in data similar to the present data.

select the best fitting model that we judge to be interpretable (i.e., friends that are important in the model's predictions make substantive sense).

Model testing. As shown in Figure 1a, we will select our candidate models based on the training data, complete an interim registration of our model selection (in a publicly-available, timestamped pre-registration on osf.io), and then test the selected models' accuracy using the (heldout) test data. To guard against overfitting, we will select one candidate model per outcome variable. In addition to our candidate models, we will also test the out-of-sample accuracy for the non-selected models as exploratory analyses, but we will clearly distinguish selected from non-selected models (which can be verified in our registration). This could include better fitting but less interpretable models, potentially providing insight into the extent to which prioritizing interpretability helps or harms out-of-sample predictive accuracy. Figure 1c shows our approach to model testing. Figure 1c highlights the independence of the model and decision making from the test data, including filtering friends (i.e., feature selection) and scoring the PCA solution (if two-step PCR is chosen). The model's performance in the test data should thus be well-guarded against overfitting and data leakage.

Outcome-Neutral Quality Checks

We will assess the self-report and Twitter data for quality using outcome neutral criteria to better ensure that we can trust our results, especially if we find low predictive accuracy.

Self-report quality. To trust our predictive accuracy results we need to ensure the quality of the self-report data. We will do this in two ways. The first will be by assessing scale reliability via split-half reliability. The second will be by ensuring there are not floor or ceiling effects in our data by plotting distributions of scale scores. If any of the four scale show low reliability or

evidence of a floor or ceiling effect, we will consider that scale as having failed this quality check and we will consider predictive modeling with that scale less informative.

Twitter friends. Because each single friend is expected to contribute at most a small amount to predictive accuracy, our analyses rest on the presence of many (different) friends in the data. As such, we will examine the number of friends left in the data under the three feature selection thresholds we use, a minimum of 3, 4, or 5 followers in the data. We will not use a feature selection threshold that results in fewer than 100 friends unless even the least strict filter (minimum of 3 followers) does. In this event, we will proceed with analyses but consider this 'minimum number of friends in the data' quality check failed, and we will note that in the interpretation of the predictive modelling.

Preliminary Feasibility Analyses

We have conducted a series of preliminary analyses aimed at determining the feasibility of our planned approach, predicting the sentiment and emotion of user's tweets from their Twitter friends. The scripts and data for these analyses can be found on OSF at the following link: https://osf.io/ky7u3/. We consider these to be feasibility analyses only; they do not have strong implications for our anticipated findings for at least three reasons. First, the self-report mental health variables we are seeking to predict likely have different psychometric properties than average sentiment; the former having been subject to more rigorous psychometric testing than the latter. Second, the sampling procedure here is very different than the sampling procedure we employed for the actual study. Third, the combined size of the initial samples is smaller by almost 200 participants and the replication sample is even smaller, leading to less precision. For these reasons, we believe these results speak only to the feasibility of the proposed analyses and should be interpreted primarily in this light.

Despite these differences, we conducted these analyses to explore some peculiarities of using a user-friend matrix to predict user characteristics, such as sparsity of the data. Indeed, this exercise provided valuable insights that informed our planned analyses. For example, we found that three or more followers in the data was a good lower-bound for feature selection; even relaxing this to two or more followers in the data led to model convergence issues. Likewise, we found that classic LASSO produced impossible solutions in this kind of data, whereas Relaxed LASSO did not suffer from such issues. Thus, these feasibility analyses provided a means to work out some of the issues inherent to using sparse, noisy predictors like Twitter friends.

We started by taking two samples of twitter users which were ultimately combined. The first came from a random sample of the first author's two-step Twitter friend network (a random sample of his friends' friends; $n_{two\text{-}step\text{-}friends}$ = 282) and the second came from a random sample of followers from two prominent political accounts ('@BarackObama' & '@realDonaldTrump'; $n_{political\text{-}followers}$ = 532). We accessed the Twitter API to download full friends lists and all available tweets for these 814 accounts. We next removed users that wouldn't have met our inclusion criteria; users were removed if their language set to anything other than English, if they had fewer than 25 friends, and if they had fewer than 25 tweets; this resulted in a final combined sample of 484 Twitter users. We split the data into a 60-40 training-test split ($n_{training}$ = 290; n_{test} = 194). We then scored each of these users' tweets for sentiment and emotion using the NRC lexicons developed for scoring sentiment and emotion in tweets (Mohammad & Kiritchenko, 2015); for the sake of space, we'll just discuss the sentiment results here.

We predicted tweet sentiment from twitter friends' using the approaches outlined above[5]. Figure 2a contains model performance during training in these data as determined by the multiple correlation (R). The feature selection filter (i.e., the minimum number of followers in the data friends must have) is on the y-axis and multiple correlation (R) units are on the x-axis (ranging from 0 to .9 at the edge of the graph). Each dot represents a single model from training and bars represent the average across models within a given approach; the approach used is denoted by the color of the dots and bars. For example, the three (tightly clustered) purple dots at the top represent the average multiple correlation for the three Unsupervised PCA + ridge models fit with the user-friend matrix trimmed to only have friends with at least 5 followers in the data; the three models represent different values for lambda (i.e., different strengths of the L2 penalty). The accompanying purple bar is the average R across the three models.

As seen in Figure 2a, Random Forests with the most inclusive friends list (minimum of 3 followers in the data) had the best fitting single model (multiple R of approximately .43); Random Forests with the least inclusive friends list (minimum of 5 followers in the data) had the best average fit. We chose the former model (Random Forests with friends that have a minimum of 3 followers in the data), given that it had the best fitting single model and nearly identical average training (see Figure 2a). The out-of-sample accuracy for this model was lower than training performance, but only very slightly (R = .42). As stated previously, we will additionally consider interpretability in the planned analyses, which was not part of the decision process in the feasibility analyses.

[5] We also used classic LASSO and found that it produced impossible predictions (RMSEs and MAEs in the millions); those can be found in the script and output for the initial samples ("predictive_modelling_both_samples.html") but are not discussed here.

As a final step, we conducted a small replication, obtaining a small sample by sampling from follower lists of 10 popular Twitter accounts ('@joelembiid', '@katyperry', '@jimmyfallon', '@billgates', '@oprah', '@kevinheart4real', '@wizkhalifa', '@adele', '@nba', and '@nfl'); after applying the same filtering criteria as above, we had a replication sample of $N_{replication}$ = 129 unique users. We split these data into a 80-20 training-test split ($n_{replication\_training}$ = 103; $n_{replication\_test}$ = 26). The training results are displayed in Figure 2b in a graph with an identical layout to Figure 2a. As seen in Figure 2b, results looked very similar in the replication data as they did in the initial samples, with Random Forests using friends with a minimum of three followers in the data again performing best (R = .61); Random Forests using friends with a minimum of 5 followers in the data again had a slightly better average performance. We selected the former model (Random Forests, friends with 3+ followers) and again found a small decrease in predictive accuracy when applying this model to the test data (R = .48), thus confirming the relative consistency of our modelling workflow.

References

Anderson, C., Keltner, D., & John, O. P. (2003). Emotional Convergence Between People over Time. *Journal of Personality and Social Psychology*, *84*(5), 1054–1068. https://doi.org/10.1037/0022-3514.84.5.1054

Back, M. D., Stopfer, J. M., Vazire, S., Gaddis, S., Schmukle, S. C., Egloff, B., & Gosling, S. D. (2010). Facebook profiles reflect actual personality, not self-idealization. *Psychological Science*, *21*(3), 372–374. https://doi.org/10.1177/0956797609360756

Bair, E., Hastie, T., Paul, D., & Tibshirani, R. (2006). Prediction by supervised principal components. *Journal of the American Statistical Association*, *101*(473), 119–137. https://doi.org/10.1198/016214505000000628

Brewin, C. R., Rose, S., Andrews, B., Green, J., Tata, P., McEvedy, C., Turner, S., …, Foa, E. B. (2002). Brief screening instrument for post-traumatic stress disorder. *The British Journal of Psychiatry*, *181*(2), 158-162. https://doi.org/10.1192/bjp.181.2.158

Buss, D. M. (1984). Toward a psychology of person-environment (PE) correlation: The role of spouse selection. *Journal of Personality and Social Psychology*, *47*(2), 361–377. https://doi.org/10.1037/0022-3514.47.2.361

Coppersmith, G. A., Harman, C. T., & Dredze, M. H. (2014). Measuring Post Traumatic Stress Disorder in Twitter. *In Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM).*, *2*(1), 23–45.

De Choudhury, M., Counts, S., & Horvitz, E. (2013). Predicting postpartum changes in emotion and behavior via social media. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*, 3267. https://doi.org/10.1145/2470654.2466447

De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media Predicting*, *2*, 128–137. https://doi.org/10.1109/IRI.2012.6302998

De Choudhury, M., Kiciman, E., Dredze, M., Coppersmith, G., & Kumar, M. (2016). Discovering Shifts to Suicidal Ideation from Mental Health Content in Social Media. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*, 2098–2110. https://doi.org/10.1145/2858036.2858207

Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A., & Danforth, C. M. (2011). Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter. *PLoS ONE*, *6*, e26752. https://doi.org/10.1371/journal.pone.0026752

Gentry, J. (2015). twitteR: R Based Twitter Client. R package version 1.1.9. https://CRAN.R-project.org/package=twitteR

Golbeck, J., Robles, C., Edmondson, M., & Turner, K. (2011). Predicting personality from twitter. *Proceedings - 2011 IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing, PASSAT/SocialCom 2011*, 149–156. https://doi.org/10.1109/PASSAT/SocialCom.2011.33

Gosling, S. D., Ko, S. J., Mannarelli, T., & Morris, M. E. (2002). A room with a cue: Personality judgments based on offices and bedrooms. *Journal of Personality and Social Psychology*, *82*(3), 379–398. https://doi.org/10.1037//0022-3514.82.3.379

Gross, J. J. (2002). Emotion regulation: Affective, cognitive, and social consequences. *Psychophysiology*, *39*, 281–291. http://doi.org/10.1017.S0048577201393198

Kosinski, M., & Stillwell, D. J. (2011). myPersonality Project. Retrieved from https://sites.google.com/michalkosinski.com/mypersonality

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(15), 5802–5805. https://doi.org/10.1073/pnas.1218772110

Kosinski, M., Wang, Y., Lakkaraju, H., & Leskovec, J. (2016). Mining big data to extract patterns and predict real-life outcomes. *Psychological Methods*, *21*(4), 493–506. https://doi.org/10.1037/met0000105

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*(29), 8788-8790. https://doi.org/10.1073/pnas.1320040111.

Kuhn, M. (2008). Caret package. Journal of Statistical Software, 28(5). https://www.jstatsoft.org/article/view/v028i05

Kuhn, M. & Johnson, K. (2013). Applied predictive modelling. New York, NY: Springer-Valeg.

Leary, M. R., & Kowalski, R. M. (1990). Impression management: A literature review and two-component model. *Psychological Bulletin*, *107*(I), 34–47. https://doi.org/10.1037/0033-2909.107.1.34

Meinshausen, N. (2007). Relaxed Lasso. *Computational Statistics and Data Analysis*, *52*(1), 374–393. https://doi.org/10.1016/j.csda.2006.12.019

Mohammad, S. M., & Kiritchenko, S. (2015). Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, *31*(2), 301–326. https://doi.org/10.1111/coin.12024

Nadeem, M., Horn, M., & Coppersmith, G. (2016). Identifying Depression on Twitter. *arXiv preprint arXiv:1607.07384*.

Park, M., Cha, C., & Cha, M. (2012). Depressive Moods of Users Portrayed in Twitter. *The ACM SIGKDD Workshop on Healthcare Informatics (HI-KDD)*, 1–8.

Park, G., Schwartz, A. H., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., …, Seligman, M. E. P. (2015). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, *108*(6), 934–952. https://doi.org/10.1037/pspp0000020

Paulhus, D. L., & Trapnell, P. D. (2008). Self-presentation: An agency-communion framework. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality psychology* (pp. 492-517). New York: Guilford.

Pew Research Center (March 2018). Social Media Use in 2018. Retrieved from http://www.pewinternet.org/2018/03/01/social-media-use-in-2018/

Pilkonis, P. A., Choi, S. W., Reise, S. P., Stover, A. M., Riley, W. T., & Cella, D. (2011). Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS®): depression, anxiety, and anger. *Assessment*, 18, 263-283.

Qiu, L., Lin, H., Ramsay, J., & Yang, F. (2012). You are what you tweet: Personality expression and perception on Twitter. *Journal of Research in Personality*, *46*(6), 710–718. https://doi.org/10.1016/j.jrp.2012.08.008

Quercia, D., Kosinski, M., Stillwell, D., & Crowcroft, J. (2011). Our twitter profiles, our selves: Predicting personality with twitter. *Proceedings - 2011 IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing, PASSAT/SocialCom 2011*, (May 2014), 180–185. https://doi.org/10.1109/PASSAT/SocialCom.2011.26

R Core Team (2018). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria.* URL https://www.R-project.org/.

Reece, A. G., Reagan, A. J., Lix, K. L. M., Dodds, P. S., Danforth, C. M., & Langer, E. J. (2017). Forecasting the onset and course of mental illness with Twitter data. *Scientific Reports*, *7*(1), 1–11. https://doi.org/10.1038/s41598-017-12961-9

Roesslein, J. (2009). Tweepy v3.6.0. http://docs.tweepy.org/en/latest/

Schaefer, D. R., Kornienko, O., & Fox, A. M. (2011). Misery does not love company: Network selection mechanisms and depression homophily. *American Sociological Review*, *76*(5), 764–785. https://doi.org/10.1177/0003122411420813

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., … Ungar, L. H. (2013). Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLoS ONE*, *8*(9), e73791. https://doi.org/10.1371/journal.pone.0073791

Srivastava, S. (2018, November 28). Sound Inference in Complicated Research: A Multi-Strategy Approach. https://doi.org/10.31234/osf.io/bwr48

Verduyn, P., Lee, D. S., Park, J., Shablack, H., Orvell, A., Bayer, J., …, Kross, E. (2015). Passive facebook usage undermines affective well-being: Experimental and longitudinal evidence. *Journal of Experimental Psychology: General*, *144*(2), 480–488. https://doi.org/10.1037/xge0000057

Watson, D., Beer, A., & Mcdade-Montez, E. (2014). The role of active assortment in spousal similarity. *Journal of Personality*, *82*(2), 116–129. https://doi.org/10.1111/jopy.12039

Watson, D., Hubbard, B., & Wiese, D. (2000a). General traits of personality and affectivity as predictors of satisfaction in intimate relationships: evidence from self- and partner-ratings. *Journal of Personality*, *68*(3), 413–449. https://doi.org/10.1111/1467-6494.00102

Watson, D., Hubbard, B., & Wiese, D. (2000b). Self-other agreement in personality and affectivity: The role of acquaintanceship, trait visibility, and assumed similarity. *Journal of Personality and Social Psychology*, *78*(3), 546–558. https://doi.org/10.1037/0022-3514.78.3.546

Watson, D., Klohnen, E. C., Casillas, A., Simms, E. N., Haig, J., & Berry, D. S. (2004). Match makers and deal breakers: Analyses of assortative mating in newlywed couples. *Journal of Personality*, *72*(5), 1029–1068. https://doi.org/10.1111/j.0022-3506.2004.00289.x

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science*, 174569161769339. https://doi.org/10.1177/1745691617693393

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, *112*(4), 1036–1040. https://doi.org/10.1073/pnas.1418680112

Youyou, W., Stillwell, D., Schwartz, H. A., & Kosinski, M. (2017). Birds of a Feather Do Flock Together: Behavior-Based Personality-Assessment Method Reveals Personality Similarity Among Couples and Friends. *Psychological Science*, *28*(3), 276–284. https://doi.org/10.1177/0956797616678187

Table 1. Sample Size from Each Recruitment Method

Recruitment Method	$N_{Prescreen}$	$N_{Eligible}$	N_{Final}
UOHSP	291	87	80
mTurk	65	39	37
Twitter Ads + reddit	955	591	505
Reddit	52	45	39
Total:	1363	762	661

UOHSP = University of Oregon Human Subjects Pool; mTurk = Amazon's mechanicalTurk; Twitter Ads + reddit = participancts recruited through promoted tweets, includes participants incidentally recruited from Reddit. Reddit = participants intentionally recruited through Reddit's R/beermoney. $N_{Prescreen}$ = all participants who completed the survey before manual screening was completed. $N_{eligible}$ = participants who met inclusion criteria. N_{Final} = all participants who met inclusion criteria and we were able to get API friends data for.

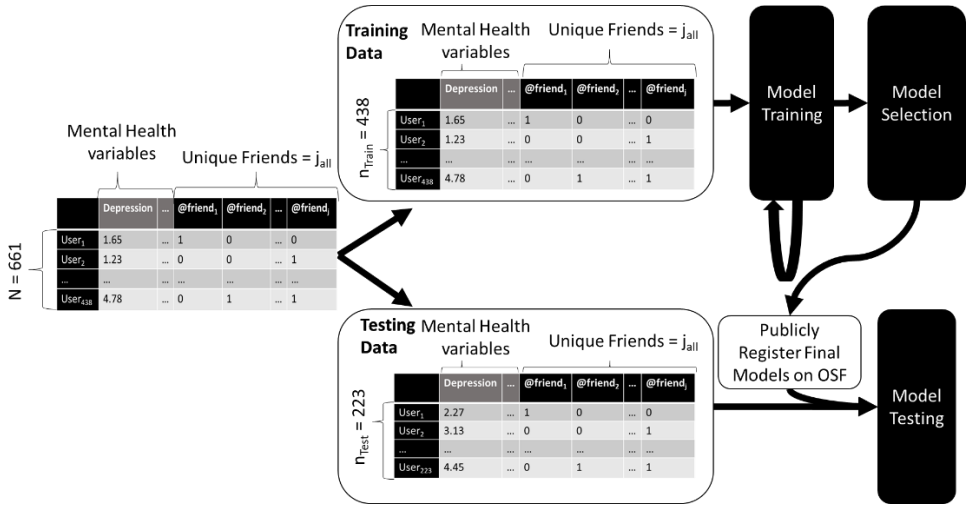

Figure 1a. Overview of Data Analysis Workflow

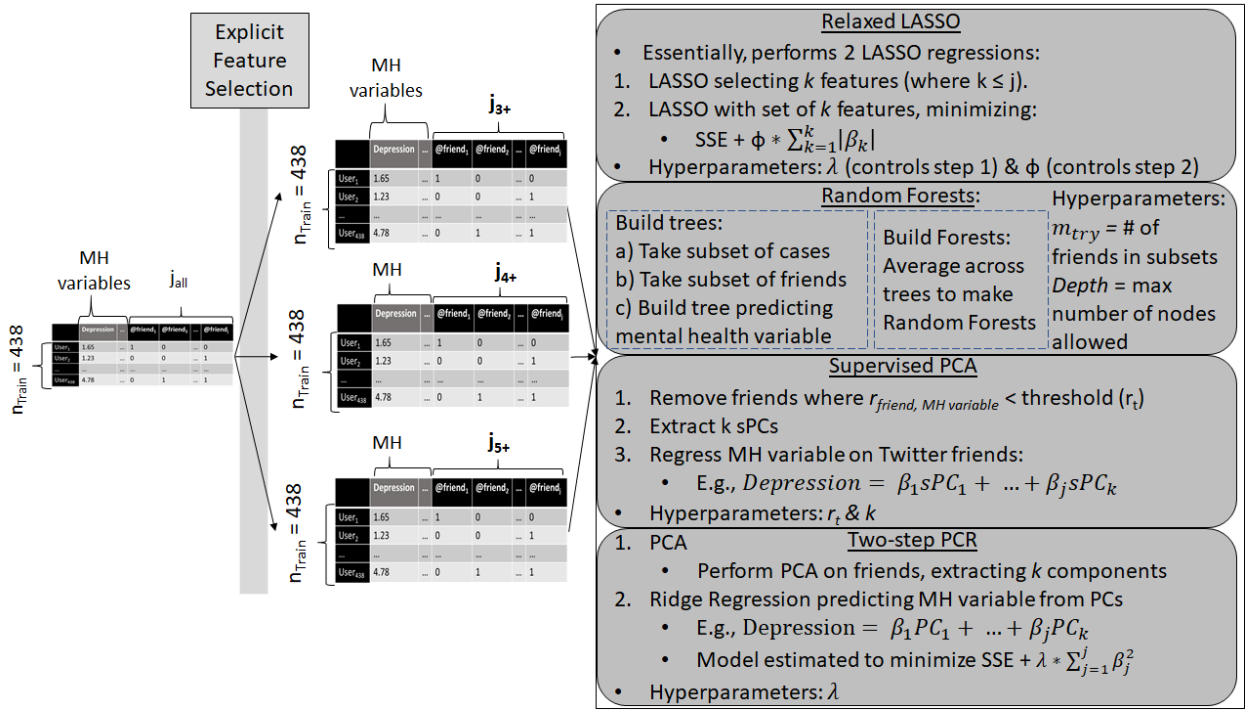

Figure 1b. Overview of Model Training Workflow with Details about Modelling Approaches.

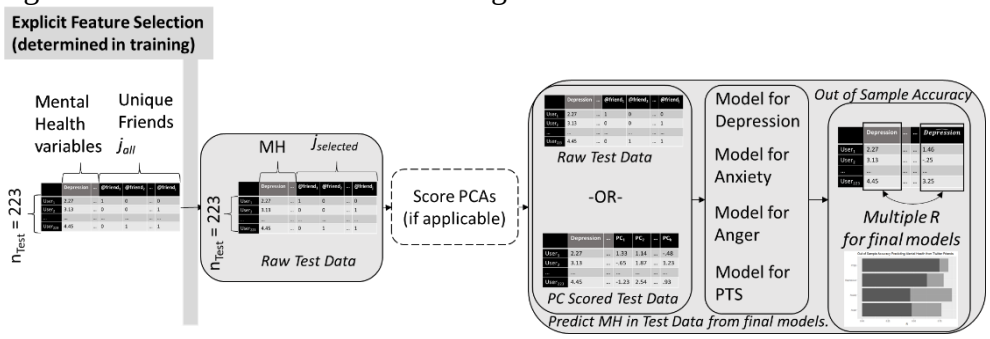

Figure 1c. Overview of Model Testing Workflow.

Note. j_{all} = number of unique friends in data. j_{3+} = number of friends with 3 or more followers in the data; j_{4+} = number of friends with 4 or more followers in the data; J_{5+} = number of friends with 5 or more followers in the data. $J_{selected}$ = friends selected for testing (based on training). MH variable = mental health variable and refers to depression, anxiety, anger, and post-traumatic stress. PCA = principal components analysis. sPCs = supervised principal components. PCs = (unsupervised) principal components. Model performance during training will be determined via k-fold cross-validation (k = 10). In Figure 1c, the dashed box is unique to two-step PCR; this would not be part of the workflow for the other three approaches.

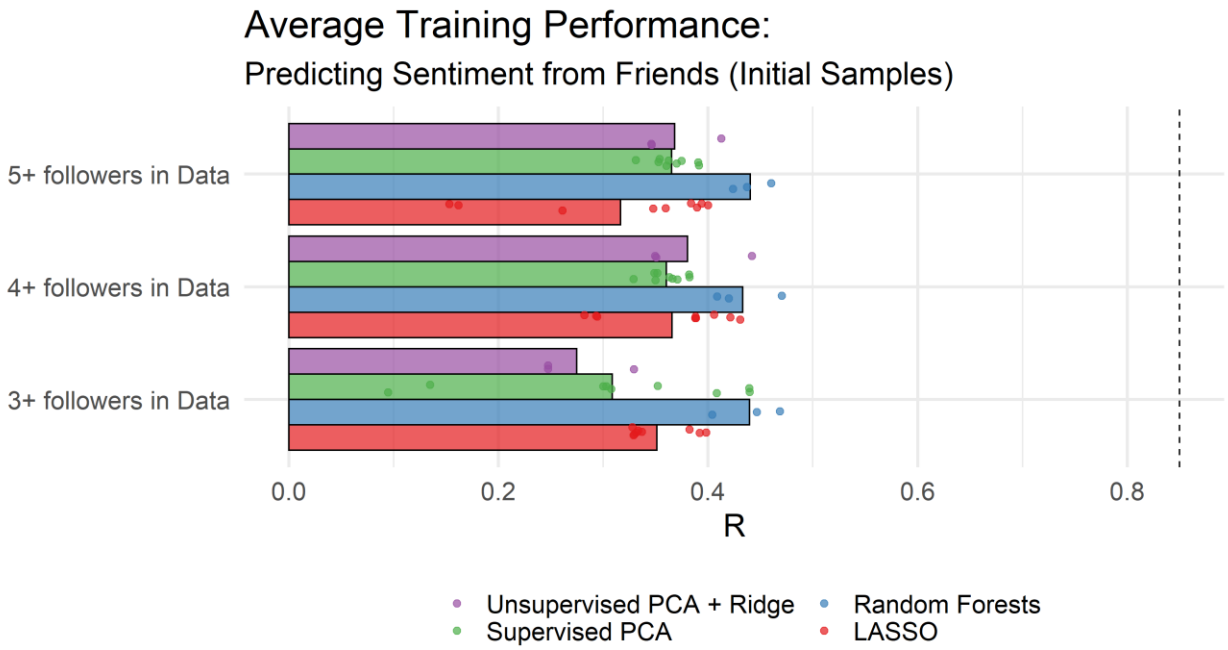

Figure 2a. Average Training Performance Predicting Tweet Sentiment from Friends in Initial (preliminary) Samples

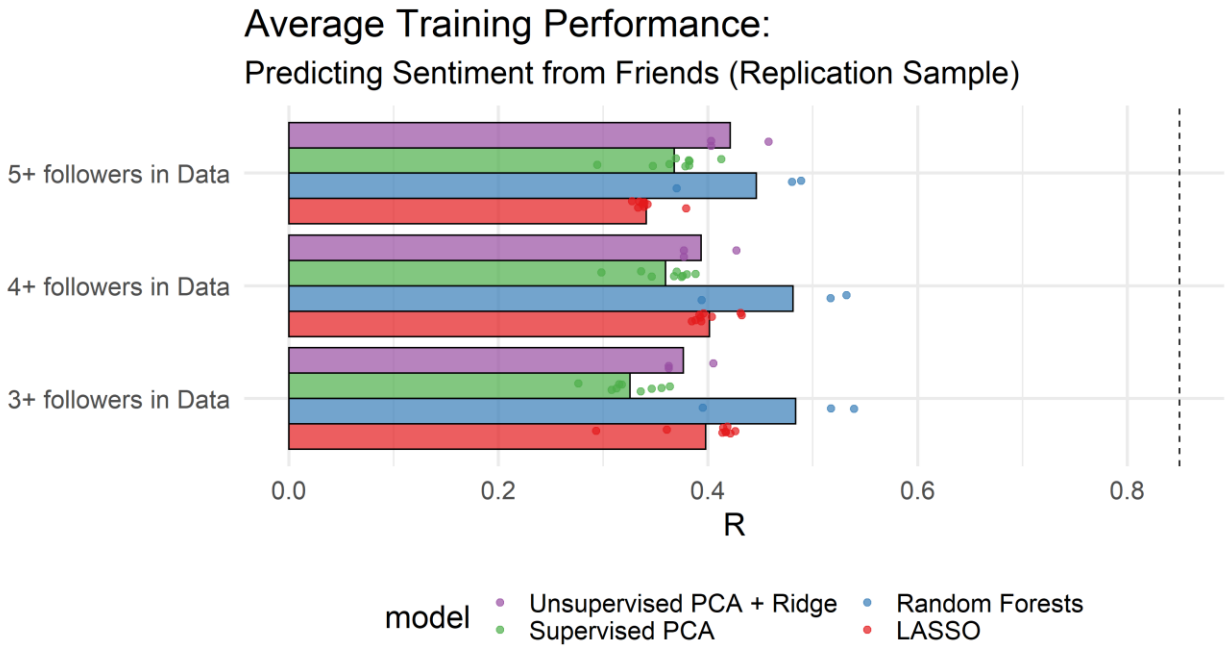

Figure 2b. Average Training Performance Predicting Tweet Sentiment from Friends.

Note. Each dot represents the average multiple correlation (averaged across k-fold runs) for a model and set of hyperparameters. The bars represent the average across training runs with different hyperparameters. The dotted line at the righthand side of the graph is the split-half reliability for tweet sentiment.