Using Matched Samples to Estimate the Effects of Exercise on Mental Health from Twitter

Virgile Landeiro Dos Reis and **Aron Culotta**
Department of Computer Science
Illinois Institute of Technology
Chicago, IL 60616
vlandeir@hawk.iit.edu, aculotta@iit.edu

Abstract

Recent work has demonstrated the value of social media monitoring for health surveillance (e.g., tracking influenza or depression rates). It is an open question whether such data can be used to make causal inferences (e.g., determining which activities lead to increased depression rates). Even in traditional, restricted domains, estimating causal effects from observational data is highly susceptible to confounding bias. In this work, we estimate the effect of exercise on mental health from Twitter, relying on statistical matching methods to reduce confounding bias. We train a text classifier to estimate the volume of a user's tweets expressing anxiety, depression, or anger, then compare two groups: those who exercise regularly (identified by their use of physical activity trackers like Nike+), and a matched control group. We find that those who exercise regularly have significantly fewer tweets expressing depression or anxiety; there is no significant difference in rates of tweets expressing anger. We additionally perform a sensitivity analysis to investigate how the many experimental design choices in such a study impact the final conclusions, including the quality of the classifier and the construction of the control group.

1 Introduction

Social media are increasingly used for tracking health concerns such as influenza (Lampos and Cristianini 2010; Culotta 2010; Paul and Dredze 2011; Signorini, Segre, and Polgreen 2011; Sadilek, Kautz, and Silenzio 2012), E. coli (Stewart and Diaz 2012), Adderall use (Hanson et al. 2013), dental pain (Heaivilin et al. 2011), insomnia (Jamison-Powell et al. 2012) and depression (De Choudhury et al. 2013). See Dredze (2012) for an overview. This approach provides an attractive complement to traditional survey approaches; it is cheaper, faster, and typically provides larger sample sizes, making it particularly appealing for monitoring diseases with rapid transmission rates.

While health tracking has been well studied, little efforts have been made to use social media data for a potentially more powerful health application: Web scale observational studies. Epidemiologists commonly conduct observational studies using survey data or health screenings to estimate causal effects when more rigorous controlled studies are infeasible or unethical — e.g., measuring the health effects of living in proximity to waste landfill sites.

Because subjects are not randomly assigned to experimental and control groups, observational studies do not have the same internal validity of randomized controlled trials, making them susceptible to the correlation/causation fallacy. That is, the observed difference in groups may be due to variables other than that proposed by the scientific hypothesis. As a result, many statistical techniques have been developed to estimate causal effects from observational data — for example, stratification and matching (Winship and Morgan 1999). These approaches all assume the presence of some observable covariates (e.g., demographics) that are predictive of group assignment.

In this paper, we take initial steps to explore the potential of Web-scale observational studies in the context of a specific health hypothesis: *Does exercise improve mood?* This question has been studied using traditional small study designs in psychology and psychiatry, where the evidence suggests that vigorous physical activity can alleviate symptoms of mild depression, improve self-image and social skills, and reduce anxiety (Taylor, Sallis, and Needle 1985). Traditional studies are typically limited by cost to small sample sizes and brief time windows. Social media provides a potential new data source to conduct such observational studies.

We develop an experimental framework to test this hypothesis using Twitter data, focusing on three research questions:

RQ1. Can we accurately annotate users according to mood and physical activity? We train a text classifier that identifies three different mood states (Hostility, Dejection, Anxiety) with 87% accuracy. Additionally, we use the increased popularity of activity tracking applications (e.g., Nike+) to identify physically active users.

RQ2. How can we identify a suitable control set of users? We provide an exact matching approach that identifies users with similar characteristics as the physically active set of users (based on gender, location, and online activity).

RQ3. How sensitive are the results to choice of control

set and quality of the classifier? We find that a more naive selection of a control group leads to inflated estimates of causal effects; additionally, we find that poor classifier accuracy can make it difficult to identify significant differences between groups.

As classifying online users by mental health or mood has been studied previously (De Choudhury, Counts, and Horvitz 2013; Li, Wang, and Hovy 2014), our primary contributions are (1) a new matching methodology for studies of online users, (2) an empirical comparison of the effect that control group selection has on the study conclusions, (3) a demonstration of this methodology for investigating the relationship between exercise and mood.

In the remainder of the paper, we first summarize related work, then describe our methodology for data collection, matching, and mood classification. We present and discuss the results, then conclude with future work.[1]

2 Related Work

Of the many recent studies inferring health from social media (Lampos and Cristianini 2010; Culotta 2010; Paul and Dredze 2011; Signorini, Segre, and Polgreen 2011; Heaivilin et al. 2011; Dredze 2012; Jamison-Powell et al. 2012; Sadilek, Kautz, and Silenzio 2012; Stewart and Diaz 2012; Hanson et al. 2013; Culotta 2013; De Choudhury et al. 2013; De Choudhury, Counts, and Horvitz 2013), perhaps the most aligned with our vision is Sadilek and Kautz (2013), who infer adverse health symptoms (coughing, headaches) from Twitter data and correlate them with environmental variables. While they do include some socio-economic control variables in a regression model, these are at the population level (zip-code), not individual covariates.

While most social media-based point prevalence work has made no attempt to control for confounding bias, Gayo-Avello (2011) emphasizes the importance of this bias, and provides evidence that age bias can affect attempts to predict political elections from Twitter sentiment. Additionally, Schonlau et al. (2009) use propensity score matching to adjust for selection bias in web surveys. Recent work has performed controlled experiments (Kohavi et al. 2009) and quasi-experiments (Oktay, Taylor, and Jensen 2010) on social media systems, though not for health studies, and not with experimental variables inferred from text.

More recently, Murnane and Counts (2014) identified Twitter users who attempted to quit smoking, then identified attributes that distinguished those who succeeded from those who did not. For example, they found that those who did not succeed posted more frequently and used less positive language, while those who succeed had greater social connectivity. This provides a great example of the promise of social media for this type of observational study, as well as the difficulty of estimating causal effects from noisy, online data.

[1] Code is available here: https://github.com/tapilab/aaai-2015-matching.

3 Methods

Below we describe our approach to sample users for the physically active (experimental) group and a control group, as well as how we classify messages according to mood.

3.1 Detecting Physically Active Users

We first identify a set of users who are physically active. Recently, a number of physical activity monitoring applications have been developed that allow one to record progress toward fitness goals. One feature of such applications is the ability to broadcast to others one's exercise activities. For example, `runmeter` allows one to broadcast the distance and time of a run.

We manually identified 10 hashtags for activity tracking applications: `runkeeper`, `nikeplus`, `runtastic`, `endomondo`, `mapmyrun`, `strava`, `cyclemeter`, `fitstats`, `mapmyfitness`, `runmeter`. We collect 123K tweets matching one of these 10 hashtags over a 10 day period (5/21/2014-5/31/2014) from 67K unique users. From these 67K users, we remove those who follow more than 5,000 accounts and those which are followed by more than 5,000 accounts. This is made in order to:

1. remove marketing accounts: physical activity tracking applications tend to do self promotion using Twitter;
2. to avoid a large number of calls to the Twitter API while collecting information about a unique user.

3.2 Matching

Below we first motivate the need for matching, then describe our procedure for identifying a matched control group.

Let H_i be a binary random variable representing a potential health outcome for individual i, e.g., $H_i = 1$ can indicate that individual i has high anxiety. In a typical observational study, individuals are categorized into two groups: the *treatment* group G_t is exposed to conditions that are hypothesized to affect the health outcome H (e.g., physical activity); the *control* group G_c is unexposed. Group assignment is indicated by a superscript: H_i^t is the health outcome for individual i in the treatment group, and H_j^c is the outcome for individual j in the control group.

The main quantity of interest is the *treatment effect*, the difference in outcome introduced by treatment: $\tau_i = H_i^t - H_i^c$. Of course, an individual i cannot be assigned to both groups simultaneously, so we can observe only one of H_i^t and H_i^c for each individual i. As a result, the average treatment effect is instead used, typically estimated by

$$\bar{\tau} = \bar{H}^t - \bar{H}^c \tag{1}$$

where $\bar{H}^t = \frac{1}{|G_t|} \sum_{i \in G_t} H_i^t$ and $\bar{H}^c = \frac{1}{|G_c|} \sum_{i \in G_c} H_i^c$. That is, the treatment effect is estimated by the difference between the mean outcomes of the two groups.

The key difference between an observational study and a randomized experiment is that in an observational study individuals are not randomly assigned to treatment and control groups. Because of this, the estimate in Equation 1 is susceptible to *omitted variable bias* when there exist confounding variables that influence both treatment and outcome. For example, if people who exercise regularly have higher income

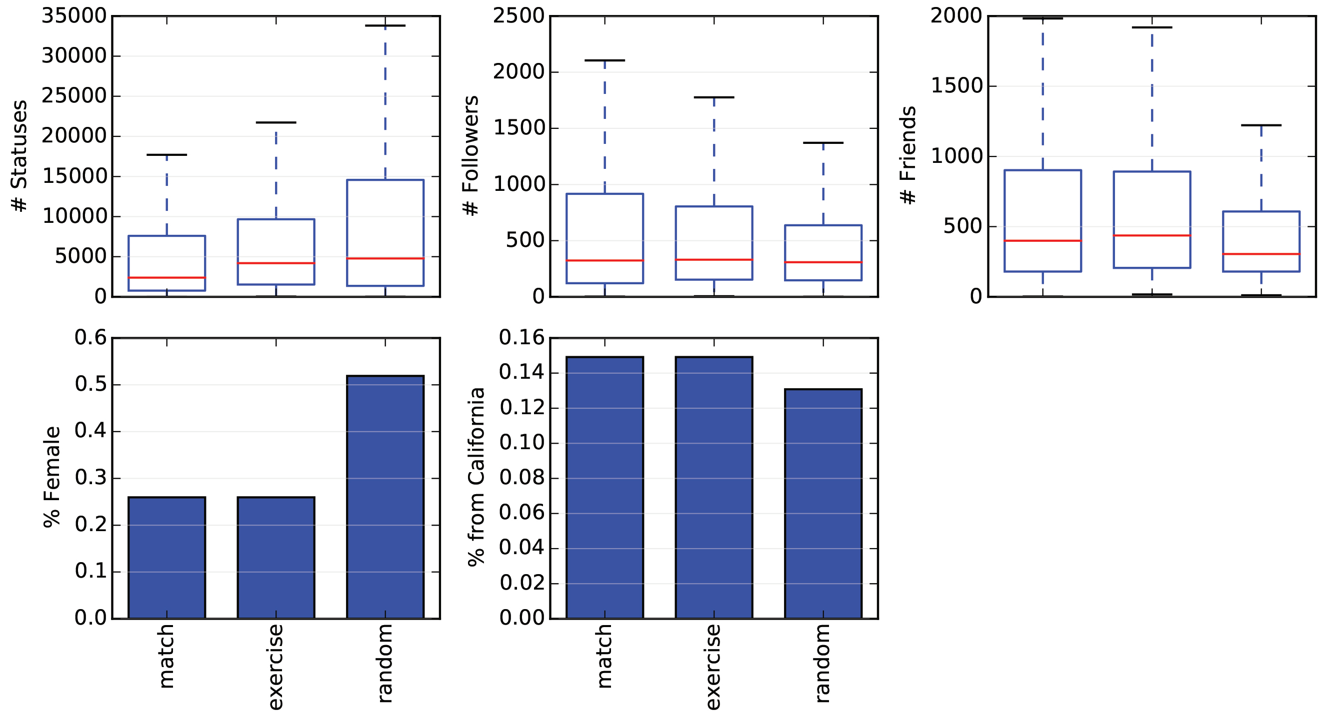

Figure 1: A comparison of various attributes of the experimental group (exercise), the matched control group, and a random control group, indicating that the matching procedure results in a control group more similar to the experimental group.

levels on average, then selecting a control group uniformly at random may over-estimate the treatment effect, since those with lower income levels may also have lower health outcomes on average. Thus, the choice of control group can dramatically impact the estimated treatment effect.

Standard approaches to reducing omitted variable bias include matching, stratification, and regression modeling (Winship and Morgan 1999). All of these methods assume that one can first measure a set of k potentially confounding variables (*covariates*) $Y_{i1} \dots Y_{ik}$ for each individual i. Common covariates used in health studies include age, education, race/ethnicity, gender, and health-related behaviors such as smoking and alcohol consumption. These covariates can then be included as control variables in a regression model or used to filter or weight samples in the control group as in stratification or matching.[2]

In this work, we use matching to identify a control group that is similar to the experimental group. For each user i in the experimental group, we search for a "virtual twin" on Twitter using the following procedure:

1. Identify the set of *friends* of i (F_i) defined as the set of accounts j such that j follows i and i also follows j.
2. Filter F_i by removing accounts that do not have the same gender as i. To determine gender, we compared the first name provided in the user profile with the U.S. Census list of names by gender[3]. Ambiguous names are removed: a name is ambiguous if it appears in both the male and female census with frequencies that differ from less than a given ϵ (we use $\epsilon = 0.1$).
3. Filter F_i to those accounts with the same city/state as i. We restrict our study to the U.S., using heuristics to parse the (potentially noisy) location field of the user's profile. If no city and state match, an account with just a matching state is accepted.
4. Rank the remaining elements of F_i according to the number of tweets, followers, and friends. We compute the cosine similarity between these values for i and for each element of F_i, after first taking the log of each value.
5. Select the element of F_i with the highest cosine similarity for inclusion in the control group.
6. Check that the selected match m for i is not using any physical activity tracker. If it appears that m uses such an application, we remove m and i respectively from the control and treatment group.

Physically active users for whom we cannot find a suitable match are removed from the pool (e.g., if we cannot infer the gender or location of u, they are removed). In the end, we select 1,161 physically active users for the experimental group and 1,161 users for the control group.

To assess the quality of this matching procedure, we compare the attributes of the control and experimental groups

[2]Note however that a 'hidden bias' can still exist when unobserved variables affect treatment and outcome simultaneously.

[3]http://www.census.gov/genealogy/www/freqnames.html

with a random selection of 236 Twitter users from the U.S. for whom we could infer gender and location. Figure 1 shows that while the number of statuses (tweets) and followers is similar between the experimental (exercise) and random groups, random users tend to have many fewer friends than either the control or experimental groups. Furthermore, the random group has over twice as many female users as the other groups, and a much different geographic distribution. For comparison, we display just the fraction of users from California, which is the most common state in the data.

As location and gender are very likely to be confounding variables, these results indicate that the process of sampling social media users can have a large impact on the final results.

3.3 Mood classification

In this section, we describe our classifier to annotate tweets by mood.

Data annotation The final variable we must infer is the mood of a user. We build a text classifier to label each tweet on three mood states: **Anger/Hostility**, **Depression/Dejection**, **Tension/Anxiety**.

Building on prior work (Bollen, Mao, and Zeng 2011), we begin with the Profile of Mood States (POMS) (McNair, Heuchert, and Shilony 2003), a psychology questionnaire commonly used to assess the mood of respondents. This contains 35 total terms corresponding to one of the three mood states above. While previous work in sentiment analysis uses keywords themselves for classification (Wilson et al. 2005), we find that the polysemy and variety of usage on Twitter makes this approach too inaccurate for our purposes. Instead, we perform supervised classification, using these 35 as a seed set to collect tweets for annotation.

For each term from POMS, we search through a sample of tweets to collect contexts in which the term appears. We next compute context vectors for each term, consisting of the frequencies of unigrams in a 3-word window from each appearance. Finally, we compute similar vectors for words in an expanded vocabulary, and identify the closest n terms to the original word, using cosine similarity. For each mood state, then, we expand the term list to 100 terms.

With these 300 terms, we then search Twitter for 10 examples of each term. We then annotate each message as expressing zero or more of the three mood states. (A second annotator labeled 100 of the tweets, resulting in 75% overlap.)

This process yields 2,367 total annotated tweets: 229 are labeled positive for the Anger/Hostility dimension, 369 for the Depression/Dejection dimension, and 381 for the Tension/Anxiety dimension.

Features We tokenize each tweet by splitting on whitespace and retaining punctuation and stop words. In addition to unigrams and bigrams, we also include features from the 2001 Linguistic Inquiry and Word Count (LIWC) lexicon (Pennebaker, Francis, and Booth 2001) and emoticons. LIWC contains 74 categories and 2,300 word patterns (which includes exact matches as well as prefixes like *awake**). This lexicon was developed over a number of years to identify categories that capture emotional and cognitive cues of interest to health, sociology, and psychology. It has been used in numerous studies (James W Pennebaker 2003), including Twitter studies (Qiu et al. 2012; Schwartz and others 2013; De Choudhury et al. 2013).

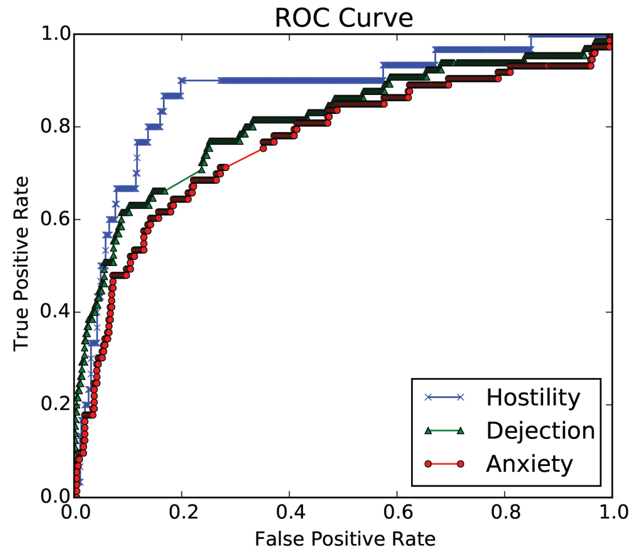

Figure 2: ROC curve for classifying tweets according to mood state.

Category	Accuracy	Precision	Recall	F1
Hostility	.901	.490	.358	.410
Dejection	.870	.620	.447	.514
Anxiety	.850	.551	.372	.442
Average	.874	.554	.392	.455

Table 1: Classification results (on tweets already matching a set of mood-related keywords).

Classifier We first perform a pilot study in which we vary the number of features from 80 to 800, comparing Decision Tree, K-neighbors, Logistic Regression, Naive Bayes, and SVM classifiers. (We use the `scikit-learn` implementations (Pedregosa and others 2011).) From this study, we select a Logistic Regression classifier, using feature selection to retain the top 160 features. Figure 2 displays the ROC curve for this classifier using 10-fold classification.

Table 1 shows the results for each category. We note that this classifier is evaluated on tweets that already contain at least one of the 300 POMS-related words identified in the previous section, so this accuracy indicates how well the classifier can disambiguate tweets that likely contain a mood state.

Table 2 shows the highest weighted coefficients for each class according to the classifier. Terms in typewriter font (`Anger` and `Sad`) indicate LIWC categories. We can see that the LIWC category is a top feature in two of the three classes. Also, bigrams appear to help here, particularly those beginning with "so" (e.g., "so tired", "so sick", "so annoy-

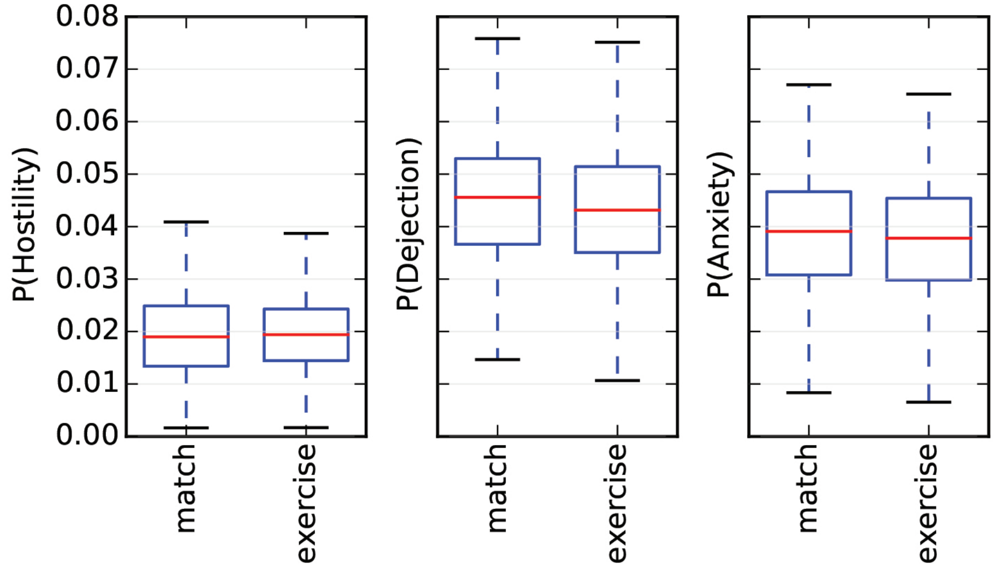

Figure 3: The results comparing the mood of users who are physically active versus a matched control group, indicating that the active group has lower rates of Dejection and Anxiety tweets. See Table 3 for significance levels.

Category	Terms
Hostility	upset, hate, so mad, fucking, irritated, shut, ive, freakin, `Anger`, your, dumb, ignorant, to die, rude, why would
Dejection	somebody, depressed, `Sad`, bored, exhausted, tired of, lazy, so tired, ive been, so sick, so depressing, get out, pain, sick, everyones
Anxiety	disgusted, shut, why cant, nervous, of this, fuck, paranoid, teacher, hate, cranky, rude, so annoying, tomorrow im, fucking, irritated

Table 2: The top 15 terms per category, ranked by fitted coefficients.

ing").

We use this classifier to annotate each user in the control and experimental groups with the level of each mood. We do this by computing the fraction of tweets from a user that are classified as positive for each class. We remove retweets as well as auto-generated tweets from consideration.

4 Results

Figure 3 shows the box-plots for the mood classifications for the experimental (exercise) and matched control groups. The y-axis is the class probability per user (the fraction of tweets assigned positive classification labels). We see that overall that Dejection and Anxiety tweets are more common than Hostility tweets. Furthermore, it appears that the control group has higher values for Dejection and Anxiety, but lower values for Hostility.

To test for significance, we use a Wilcoxon signed-rank test. Table 3 displays the results. This table reports the estimated difference in mood between the exercise and control groups. E.g., the -2.7% value in column 2 means that physically active Twitter users posted on average 2.7% fewer anxious tweets than a matched user (a relative difference). The Random Control and 50% Training Data columns use a different control set and less accurate classifier (discussed below). Bold values are those found to be statistically significant by the Wilcoxon signed-rank test ($p < 0.05$).

The "Matched Control" results in Table 3 indicate that the experimental group has a significantly lower incidence of Dejection and Anxiety tweets; the difference in Hostility tweets is not significant. These results are further validated by the psychology literature, which in small studies has found reduced incidence of anxiety and depression symptoms in highly active individuals. While it is difficult to directly compare our results to the psychology literature, for reference, exercise has been found to reduce the rate of self-reported anxiety (roughly 1 point on a 5-point scale) (Taylor, Sallis, and Needle 1985). It is suspected, however, that these differences are temporary (possibly due to endorphins). An attractive property of studies of social media is that it may be easier to study such effects over time, which we leave for future work.

4.1 Effects of control group

Table 3 also reports the results comparing the experimental group to the randomly selected control group (Random Control). We can see that such a comparison greatly overstates the effects — the experimental group has significantly lower rates of all mood classes, and the magnitude of the difference is 2-20 times that for the matched control group.

	Matched Control		Random Control		50% Training Data	
Category	**% Change**	**p-value**	**% Change**	**p-value**	**% Change**	**p-value**
Hostility	0.9	0.53	**-21.1**	9.2E-17	1.3	0.47
Dejection	**-3.9**	7.8E-4	**-5.4**	1E-4	**-2.0**	0.016
Anxiety	**-2.7**	0.02	**-7.9**	3.6E-6	0.1	0.64

Table 3: Estimated effect sizes of exercise on mood, using a Wilcoxon signed-rank test. The matched control provides a much more conservative estimate as compared to using a random control set. The results can also be sensitive to the accuracy of the classifier — using only half the training data removes the effect for Anxiety.

These results have implications for other studies that use social media data to make inferences about differences in population groups. It is important that we adjust for the selection bias introduced by the creation of the control and experimental groups. For example, Figure 1 indicates that the random control group has roughly twice as many female users as the exercise group. Given the observed lexical variation between genders on social media (Bamman, Eisenstein, and Schnoebelen 2014), this will confound hypothesis testing, since many factors other than exercise are contributing to the observed difference in mood. By matching based on gender and the other attributes described in Section 3.2, we can reduce the impact of such confounders.

4.2 Effects of classifier accuracy

Finally, Table 3 reports the results comparing the experimental group to the matched control group, but instead using a classifier trained on only 50% of the annotated data. The purpose of this sensitivity analysis was to determine how classifier quality affected the final results. We can see that effects for two of the three classes actually became stronger (Hostility and Detection), while the effect for Anxiety grew weaker (and changed sign). Taken together, this sensitivity analysis highlights the importance of the experimental design choices in conducting such studies.

5 Conclusion and Future Work

We have presented an observational study of the effect of exercise on mood using only social media data. The results of our analysis suggest that there is a smaller incidence of tweets classified as Dejection or Anxiety for users who are physically active, as determined by their use of physical activity trackers. We have also performed a sensitivity analysis which reveals the importance of selecting a realistic control group.

There are a number of limitations, most notably the fact that there is imperfection both in the assignment of users to control and experimental groups (e.g., users who exercise but do not use one of the 10 tracking apps may end up in the control group) and in the outcome variable (i.e., the mood classifier). As such imperfections are inherent to any attempt at Web observational studies, in future work we will investigate multiple imputation methods (Schafer 1999) to incorporate this uncertainty directly into hypothesis testing. Additionally, in future work we will compare alternative matching strategies, such as propensity score matching (Winship and Morgan 1999).

References

Bamman, D.; Eisenstein, J.; and Schnoebelen, T. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics* 18(2):135–160.

Bollen, J.; Mao, H.; and Zeng, X. 2011. Twitter mood predicts the stock market. *Journal of Computational Science* 2(1):1–8.

Culotta, A. 2010. Towards detecting influenza epidemics by analyzing Twitter messages. In *Workshop on Social Media Analytics at the 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Culotta, A. 2013. Lightweight methods to estimate influenza rates and alcohol sales volume from twitter messages. *Lang. Resour. Eval.* 47(1):217238.

De Choudhury, M.; Gamon, M.; Counts, S.; and Horvitz, E. 2013. Predicting depression via social media. In *ICWSM*.

De Choudhury, M.; Counts, S.; and Horvitz, E. 2013. Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, 32673276. New York, NY, USA: ACM.

Dredze, M. 2012. How social media will change public health. *IEEE Intelligent Systems* 27(4):81–84.

Gayo-Avello, D. 2011. Don't turn social media into another 'Literary digest' poll. *Commun. ACM* 54(10):121128.

Hanson, C. L.; Burton, S. H.; Giraud-Carrier, C.; West, J. H.; Barnes, M. D.; and Hansen, B. 2013. Tweaking and tweeting: Exploring twitter for nonmedical use of a psychostimulant drug (adderall) among college students. *Journal of Medical Internet Research* 15(4):e62.

Heaivilin, N.; Gerbert, B.; Page, J.; and Gibbs, J. 2011. Public health surveillance of dental pain via twitter. *Journal of Dental Research* 90(9):1047–1051. PMID: 21768306 PMCID: PMC3169887.

James W Pennebaker, M. R. M. 2003. Psychological aspects of natural language. use: our words, our selves. *Annual review of psychology* 54:547–77.

Jamison-Powell, S.; Linehan, C.; Daley, L.; Garbett, A.; and Lawson, S. 2012. "I can't get no sleep": Discussing #insomnia on Twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, 15011510. New York, NY, USA: ACM.

Kohavi, R.; Longbotham, R.; Sommerfield, D.; and Henne, R. M. 2009. Controlled experiments on the web: survey and practical guide. *Data Mining and Knowledge Discovery* 18(1):140–181.

Lampos, V., and Cristianini, N. 2010. Tracking the flu pandemic by monitoring the social web. In *2nd IAPR Workshop on Cognitive Information Processing (CIP 2010)*, 411–416.

Li, J.; Wang, X.; and Hovy, E. 2014. What a nasty day: Exploring mood-weather relationship from twitter. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, 1309–1318. ACM.

McNair, D.; Heuchert, J.; and Shilony, E. 2003. Research with the profile of mood states (poms) 1964–2002: A comprehensive bibliography. *MHC, Toronto* 7–8.

Murnane, E. L., and Counts, S. 2014. Unraveling abstinence and relapse: smoking cessation reflected in social media. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, 1345–1354. ACM.

Oktay, H.; Taylor, B. J.; and Jensen, D. D. 2010. Causal discovery in social media using quasi-experimental designs. In *Proceedings of the First Workshop on Social Media Analytics*, 1–9. ACM.

Paul, M. J., and Dredze, M. 2011. You are what you tweet: Analyzing twitter for public health. In *Fifth International AAAI Conference on Weblogs and Social Media*.

Pedregosa, F., et al. 2011. Scikit-learn: Machine learning in Python. *Machine Learning Research* 12:2825–2830.

Pennebaker, J.; Francis, J.; and Booth, R. 2001. Linguistic inquiry and word count: LIWC 2001. *World Journal of the International Linguistic Association*.

Qiu, L.; Lin, H.; Ramsay, J.; and Yang, F. 2012. You are what you tweet: Personality expression and perception on twitter. *Journal of Research in Personality* 46(6):710–718.

Sadilek, A., and Kautz, H. 2013. Modeling the impact of lifestyle on health at scale. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, 637646. New York, NY, USA: ACM.

Sadilek, A.; Kautz, H.; and Silenzio, V. 2012. Predicting disease transmission from geo-tagged micro-blog data. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.

Schafer, J. L. 1999. Multiple imputation: a primer. *Statistical methods in medical research* 8(1):3–15.

Schonlau, M.; Soest, A. v.; Kapteyn, A.; and Couper, M. 2009. Selection bias in web surveys and the use of propensity scores. *Sociological Methods & Research* 37(3):291–318.

Schwartz, H. A., et al. 2013. Characterizing geographic variation in well-being using tweets. In *Seventh International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Signorini, A.; Segre, A. M.; and Polgreen, P. M. 2011. The use of twitter to track levels of disease activity and public concern in the U.S. during the influenza a H1N1 pandemic. *PLoS ONE* 6(5):e19467.

Stewart, A., and Diaz, E. 2012. Epidemic intelligence: For the crowd, by the crowd. In Brambilla, M.; Tokuda, T.; and Tolksdorf, R., eds., *Web Engineering*, number 7387 in Lecture Notes in Computer Science. Springer Berlin Heidelberg. 504–505.

Taylor, C. B.; Sallis, J. F.; and Needle, R. 1985. The relation of physical activity and exercise to mental health. *Public health reports* 100(2):195.

Wilson, T.; Hoffmann, P.; Somasundaran, S.; Kessler, J.; Wiebe, J.; Choi, Y.; Cardie, C.; Riloff, E.; and Patwardhan, S. 2005. Opinionfinder: A system for subjectivity analysis. In *Proceedings of hlt/emnlp on interactive demonstrations*, 34–35. Association for Computational Linguistics.

Winship, C., and Morgan, S. L. 1999. The estimation of causal effects from observational data. *Annual Review of Sociology* 25:pp. 659–706.