Reducing Sampling Bias in Social Media Data for County Health Inference

Aron Culotta*

Abstract

A number of recent studies have demonstrated the utility of social media data for inferring societal attributes such as public opinion and health. A commonly declared limitation of this methodology is the selection bias inherent in this approach – social media users are a non-representative sample of the population. This is exacerbated by filtering steps that further limit the sample set in biased ways. Building on recent work in computational linguistics that infers demographic attributes of people based on their communications, we investigate methods to quantify and control for selection bias in social media studies. We present results estimating several county-level health statistics (e.g., obesity, diabetes, access to healthy foods) based on the Twitter activity of the top 100 counties in the U.S., and we compare strategies for reducing selection bias.

Key Words: social media, health, natural language processing, reweighting

1. Introduction

Social media are increasingly used for tracking health concerns such as influenza (Lampos and Cristianini, 2010; Culotta, 2010; Paul and Dredze, 2011a; Signorini et al., 2011; Sadilek et al., 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. These data provide an attractive complement to traditional survey approaches; data collection is cheaper and faster, and sample sizes are typically larger, making it particularly appealing for monitoring diseases with rapid transmission rates.

A fundamental limitation of using social media for public health applications is sampling bias. For example, Twitter users are not a representative sample of the population, tending to skew towards young, urban, minority individuals (Mislove et al., 2011). Gayo-Avello (2011) show that age bias can affect attempts to predict political elections from Twitter sentiment. Despite wide acknowledgment of this issue, there has been little done to address it.

In recent work (Culotta, 2014), we found that several statistics of county health could be estimated based on lexical patterns in geolocated Twitter messages. For example, counties that use more positive emotional terms ("happy", "best") tend to report greater socio-emotional support on government surveys; counties that use more profanity and more frequently discuss sports and television tend to have higher obesity rates. When compared to county-level models using demographics alone, including Twitter-derived variables reduced predictive error for 20 of 27 health-related statistics.

In this paper, we extend our prior work by adjusting for the demographic bias of Twitter data. We first automatically infer the race and gender of users in our sample, then compare the makeup of each county's Twitter users with U.S. Census demographics. We then use standard survey reweighting to adjust model predictions. We find that this approach reduces held-out prediction error by 4.3% on average, providing improvements for 20 of the 27 health statistics we predict. We conclude with a discussion of the implications this has for the potential of social media data for use in public health applications.

*Department of Computer Science, Illinois Institute of Technology, 3300 South Federal Street, Chicago, IL 60616

2. Related Work

We first briefly review related work in the study of language, health, and social media.

2.1 Language and Heath

Language has long been investigated as an indicator of health. For example, Gottschalk (Gottschalk and Gleser, 1979) performed a content analysis of patients to determine psychological state, such as anxiety, hostility, and alienation. Pennebaker (James W Pennebaker, 2003) provides an excellent review of research connecting linguistic patterns to demographics, personality, psychology, mental health.

While many studies support the connection between mental health and language, the connection between physical health and language is less well-established. Some studies have reported correlations between "Type A" language and heart diseases(Graham et al., 1989) and positive emotional language with longevity (Danner et al., 2001). Given growing evidence supporting the link between emotional well-being and health (Howell et al., 2007), estimating psychological health may serve as a predictive surrogate for physical health.

The emerging study of the economics of language has also investigated how language relates to decision-making, which in turn can affect health. For example, in a study of 76 countries, Chen (Chen, 2013) found that certain grammatical properties correlate with higher rates of savings and lower rates of smoking and obesity, concluding that some linguistic constructs may foster future-oriented behavior. Chiswick (Chiswick and Miller, 2007) investigates how language proficiency of immigrants can impact employment and other socio-economic factors.

2.2 Social Media and Health

There is a growing body of work investigating social media to track health concerns such as influenza (Lampos et al., 2010; Culotta, 2010; Paul and Dredze, 2011a; Signorini et al., 2011; Sadilek et al., 2012), E. coli (Stewart and Diaz, 2012), alcohol consumption (Culotta, 2013), Adderall use (Hanson et al., 2013), insomnia (Jamison-Powell et al., 2012) and depression (De Choudhury et al., 2013). Most of these focus on detecting explicit mentions of a symptom of interest (e.g., "Staying home from work today with a sore throat"). In contrast, the present work investigates more nuanced linguistic cues that correlate with the overall health of a population.

Ghosh & Guha (Ghosh and Guha, 2013) identified geo-spatial patterns in specific obesity-related tweets (e.g. "fast food"), using topic models to qualitatively characterize discussions of obesity on Twitter. While some ancillary data is used for comparison (e.g., location of fast food restaurants), no correlation analysis is performed with obesity statistics. Additionally, Paul & Dredze (Paul and Dredze, 2011b) use a topic model to discover obesity-related tweets, finding a .28 correlation with state obesity statistics.

Our methodology is most similar to that of Schwartz et al. (Schwartz et al., 2013), who find tweets to be predictive of county-level surveys of life satisfaction. Here, we also use LIWC and PERMA lexicons as features in a regression model of county statistics.

There have been few attempts made to address the selection bias in social media. 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 et al., 2010) on social media systems, though not for health studies, and not with experimental variables inferred from text.

Table 1: The 27 statistics collected for 100 counties. These are the dependent variables in our regression models. For a full description of each statistic, please consult the County Health Rankings and Roadmaps Project at http://www.countyhealthrankings.org/.

Outcomes	Behaviors	Care	Environment
Poor Health	Smoking	Ambulatory Care	Education
Unhealthy Days	Inactivity	Uninsured	Graduation Rate
Mental Health	Drinking	Primary Care	Unemployment
Low Birthweight	Driving Deaths	Dentists	Child Poverty
Diabetes	STIs	Mammography	Social Support
Obesity	Teen Birth Rate		Single Parent
			Violent Crime
			Recreational Facilities
			Access to Healthy Foods
			Fast Food

3. Data

We begin by summarizing our prior work (Culotta, 2014) that estimates county health statistics from Twitter. First, we describe how we collected the health and Twitter data and provide descriptive statistics of their contents.

3.1 County Health Data

Using data from the U.S. Census' State-Based Counties Gazetteer,[1] we collected the top 100 most populous counties in the U.S. along with their geographical coordinates. Each county is assigned a Federal Information Processing Standards (FIPS) code as a unique identifier. The County Health Rankings & Roadmaps,[2] a partnership between the Robert Wood Johnson Foundation and the University of Wisconsin Population Health Institute, aggregates county-level health factors from a wide range of sources, including the Behavioral Risk Factor Surveillance System, American Community Survey, and the National Center for Health Statistics, collected over the past three years.[3] These publicly available data contain county statistics on 30 measures of mortality, morbidity, health behaviors, clinical care, socio-economic factors, and physical environment.

For each of the top 100 most populous counties, we collected 27 health statistics (3 were removed because of missing values for some counties). These are listed in Table 1. As space precludes a precise definition of how each statistic was computed, we refer the reader to the County Health Rankings website for more information.

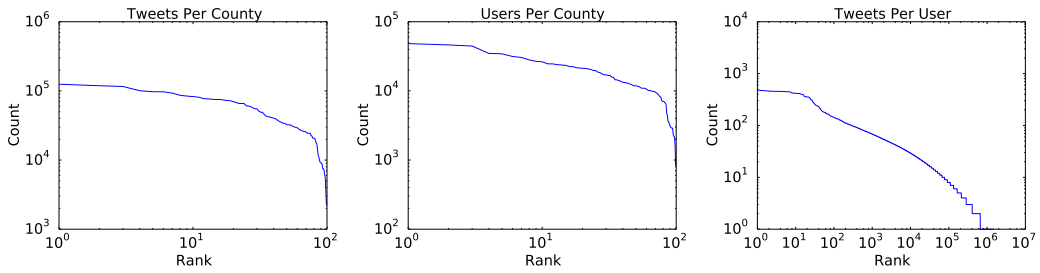
3.2 Twitter Data

We next constructed a set of 100 Twitter queries consisting of one geographical bounding box for each county, approximated by a 50 square mile area centered at the county coor-

[1] `http://www.census.gov/geo/maps-data/data/docs/gazetteer/Gaz_counties_national.zip`

[2] `http://www.countyhealthrankings.org/`

[3] While the Twitter was collected more recently, most county-level statistics, and particularly their relative differences, are slow to change.



(a) (b) (c)

Figure 1: Distributions over the 4.31M tweets, 1.46M users, and 100 counties in the dataset.

dinates obtained from the U.S. Census.[4] We then submitted these queries continuously to Twitter's search API from December 5, 2012 to August 31, 2013 (with intermittent stoppages for technical difficulties). These queries return tweets that have been geolocated, typically tweets issued from a mobile device. This resulted in 4.31M tweets from 1.46M unique users. For each tweet, we retain the tweet content as well as the user description field, a short, user-provided summary (e.g., "motivated law student"). Figure 1 shows distributions of tweets per county, users per county, and tweets per user. While the demographic distributions of Twitter users are thought to skew young and urban (Duggan and Brenner, 2013a), it is worth noting that these 1.46M users represent over 1% of the total population of these 100 counties (130M). As expected, Twitter usage varies significantly by county size. On average, we collect 14.5K users per county, with 66 counties containing at least 10K users. Hudson County (part of the New York metropolitan area) has the most with 52K users, Honolulu County the least with 845. The tweets per user graph exhibits a typical long tail — a few users tweet very often, but most tweet infrequently.

We note that this data collection methodology differs from that of Schwartz et al. (Schwartz et al., 2013), who collect the 10% "garden hose" sample of the entire Twitter stream, then use heuristics to filter by location using the user's profile information. This can yield more tweets (since only a small percentage of tweets are geocoded), but can introduce additional geolocation noise due to the unreliability of the location field (Hecht et al., 2011).

4. Linguistic Representation

Given a collection of tweets categorized by county, we next must distill them into a set of variables to correlate with the health statistics. Due to the small number of validation points (100 counties) and the large number of potential variables (hundreds of thousands of unique words), rather than considering words as variables, we instead consider word categories. We build on prior work that considers two lexicons:

- **LIWC:** The 2001 Linguistic Inquiry and Word Count lexicon (Pennebaker et al., 2001) contains 74 categories and 2,300 word patterns (which includes exact matches as well as prefixes like *awake**). Each word pattern may belong to multiple categories (e.g., *Physical*, *Sleep*). 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),

[4]This introduces a small amount of noise – 957 tweets came from overlapping bounding boxes. This can be eliminated by using the county polygon data from the Census. We thank the anonymous reviewer for this suggestion.

including Twitter studies (Qiu et al., 2012; Schwartz et al., 2013; De Choudhury et al., 2013).

- **PERMA:** The PERMA lexicon (Seligman, 2011) contains 10 categories and 1,522 words. The categories reflect the five dimensions of positive psychology (Positive emotion, Engagement, Relationships, Meaning, Achievement) — each category is either positive or negative. For example, *R+* indicates positive relationships and *P-* indicates negative emotions. Only exact matches are considered, and each word belongs to exactly one category.

We select these lexicons based on their use in prior work (Schwartz et al., 2013) and the fact that they were designed to represent categories of relevance to health and personality.

For each county, then, we record the frequency with which each lexical category is used. To do this, we use a simple tokenizer to process each tweet that removes punctuation and then splits by whitespace to return a list of tokens. Additionally, we remove all mentions and URLs. The remaining tokens are matched against the above lexicons, resulting in a vector of category frequencies for each county.

We distinguish between tokens appearing in the tweet text and tokens appearing in the user description, denoted by the prefixes (*d=*) and (*t=*). For example, [*d=Sleep: 2*, *t=R+: 1*] indicates that two tokens in the description field map to the *Sleep* category and that one token in the tweet text maps to the positive relationship category.

We found that only 70 of the LIWC categories appear in our data, along with all 10 of the PERMA categories, yielding a total of 80 linguistic categories.

For each county, we create a vector of 160 values reflecting the frequency of each category (80 categories each for description and text tokens). Since the magnitude of these values will vary greatly based on the number of tweets collected from each county, we normalize by user. That is, we store the proportion of users from the county who use a word from each category. Note that if one user tweets the same word category many times, this will only increase the numerator by one; the denominator is the total number of users from that county.

5. Experiments

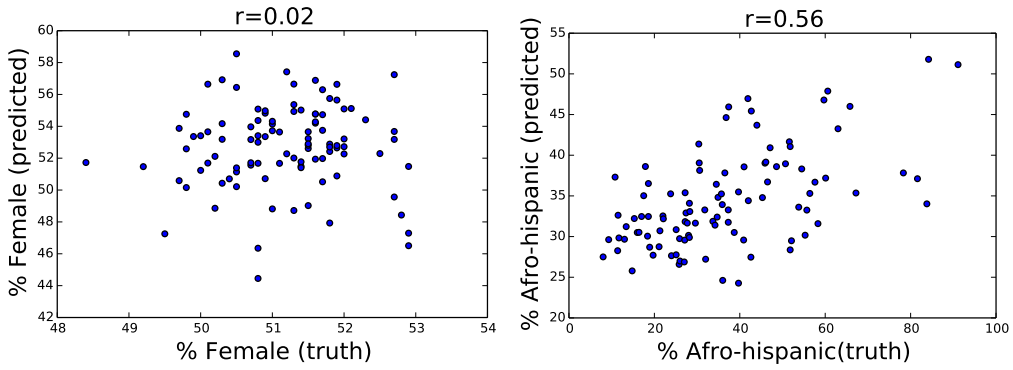
We perform regression to predict each of the 27 health-related statistics using the 180 linguistic variables described above. Given the large number of independent variables (180) relative to the number of validation points (100 counties), we use ridge regression to reduce overfitting.[5]

To estimate generalization accuracy, we use five fold cross-validation — each fold fits the model on 80 counties and predicts on the remaining 20. The splits are created uniformly at random, except that we additionally ensure that counties from the same state do not appear in both the training and test split in one fold. This is to confirm that the model is learning more than simply the state identity of each county.[6]

We use as our error metric Symmetric Mean Absolute Percentage Error (SMAPE) (Flores, 1986). SMAPE measures the relative error between the predicted and true value. This is a useful alternative to the more common mean-squared error as it can compare outcome variables that have different ranges. If y_i is the true value and $\hat{y}_i$ is the predicted value, then $\text{SMAPE} = \frac{\sum_i |y_i - \hat{y}_i|}{\sum_i y_i + \hat{y}_i}$. For non-negative y values, $\text{SMAPE} \in [0, 1]$; smaller is better.

[5] We use the implementation in `scikit-learn` (Pedregosa et al., 2011) with smoothing parameter $\alpha = 0.1$.

[6] Indeed, we find that splitting at random instead of by state increases the overall average correlation for the LIWC model from .25 to .29.



(a) (b)

Figure 2: Comparison of the estimate demographics per county with those reported by the U.S. Census. We can

6. Reweighting

We attempt to adjust for the sampling bias of Twitter data by reweighting based on gender and race. To do this, we first estimate the demographics of the Twitter users in our data, then reweight each user based on the mismatch between the Twitter demographics and those reported by the U.S. Census.

6.1 Inferring Gender

We infer gender from the user's first name. We examine the name field of each Twitter profile and compare the first name with the U.S. Census list of names by gender.[7] We remove ambiguous names from the list, which we define as those whose relative frequency is not at least 10% greater in one gender than the other.

Using this approach, we infer the gender of 48% of the users in our dataset. (The remainder cannot be inferred either because the name is ambiguous, not on the list, or the user has not entered a name in their profile). Of those for which we infer a gender, we find 44% to be male, and 56% to be female. This differs somewhat from recent surveys which have found the gender of Twitter users to be roughly equal (Duggan and Brenner, 2013b). There are two possible explanations for this: (1) our data is restricted to tweets that carry geocoordinates, which is a small sample of the overall Twitter population; (2) those who tweet more frequently are more likely to be in our dataset. Indeed, web traffic data suggests that 57% of visitors to Twitter are female,[8] which closely matches our estimates.

We next estimate the gender composition per county, considering only those users for which we can infer gender. We then compare these estimates to the latest U.S. Census data. Figure 2a shows a scatter plot of these values. We can see that the Twitter data greatly overestimate the female population of most counties, and that the two values are uncorrelated. For a number of counties, this mismatch is large — for one county, the estimated proportion of females is 59%, while the Census data report roughly 50% female.

6.2 Inferring Race

To infer race, we trained a text classifier based on a hand-labeled set of users. The classifier predicts the race of a user based on the description field of the user's profile.

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

[8] http://www.briansolis.com/2009/10/revealing-the-people-defining-social-networks/

Table 2: Race classification results, using 10-fold cross validation on 744 annotated users, using only terms from the user's description field.

	Precision	Recall	F1
African-American	.53	.47	.50
Hispanic	.81	.51	.62
White	.56	.72	.63
Average (weighted)	.60	.59	.58

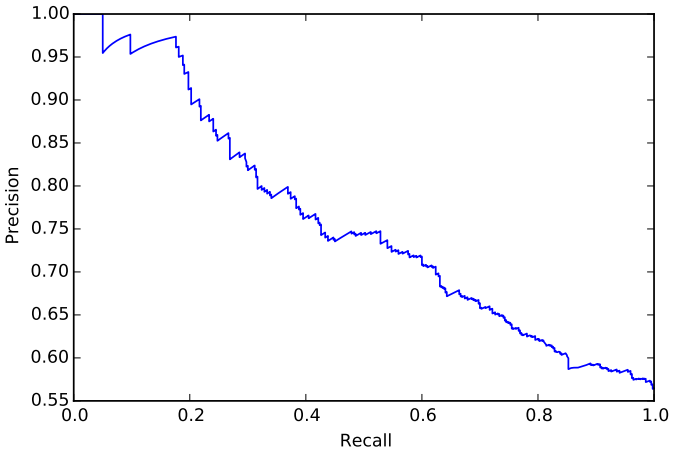


Figure 3: Precision-recall curve for race classification, averaged over three classes..

The labeled data were collected as follows: First, we used the Twitter Streaming API to obtain a random sample of users, filtered to the United States (using time zone and the place country code from the profile). From six days' worth of data (December 6-12, 2013), we sampled 1,000 profiles at random and manually categorized them by analyzing the profile, tweets, and profile image for each user. We categorized 770 Twitter profiles into one of four ethnicities (Asian, African American, Hispanic, Caucasian). Those for which ethnicity could not be determined were discarded (230/1,000; 23%).[9] The category frequency is Asian (22), African American (263), Hispanic (158), Caucasian (327). To estimate inter-annotator agreement, a second annotator sampled and categorized 120 users. Among users for which both annotators selected one of the four categories, 74/76 labels agreed (97%). There was some disagreement over when the category could be determined: for 21/120 labels (17.5%), one annotator indicated the category could not be determined, while the other selected a category. We used a variant of this data in prior work (Mohammady and Culotta, 2014).

We removed accounts that had empty description fields; we also removed the Asian users due to the small sample size, leaving a total of 744 users. Each user's description field was tokenized and converted to a binary word vector. This vector was then transformed using *tf-idf*. We then fit a logistic regression classifier with L2 regularization.

Table 2 reports the average held-out precision, recall, and F1 for the three races using 10-fold cross-validation. We see that overall precision is highest for Hispanic users, most likely because the use of Spanish makes such users easier to predict.

[9]This introduces some bias towards accounts with identifiable ethnicity; we leave an investigation of this for future work.

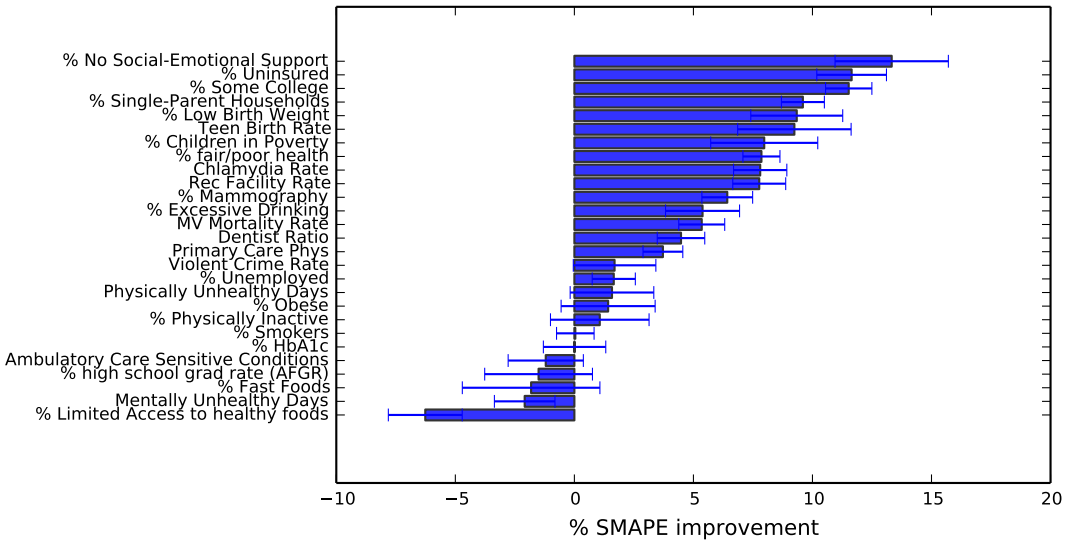


Figure 4: Reweighting based on inferred demographics reduces held-out error by 4.3% on average, improving estimate for 20 of 27 variables.

We note that these results are lower than in our previous work (Mohammady and Culotta, 2014), which used 200 tweets from each user, in addition to the user profile field we use here. The user profile is typically one sentence, so it is not surprising that accuracy is lower here. For purposes of this study, it would be difficult to collect hundreds of tweets per user. There are over 1 million users in our data, and collecting tweets for each user would require 1 million requests to the Twitter API, which has a rate limit of 15 per minute. In future work, we will consider ways to address this using the Streaming API (instead of the REST API).

To avoid injecting too much noise into the race predictions, we take a conservative approach by restricting our predictions to those for which the classifier's confidence is greater than 0.5. Figure 3 shows the precision recall curve, computed by considering different threshold values. A threshold of 0.5 corresponds to a precision of .7 and a recall of .6.

With this threshold, we assign a race to 32% of the users in our data, finding 60% White, 32% African-American, and 8% Hispanic. This over-representation of African-American users is consistent with surveys of Internet users (Duggan and Brenner, 2013b). As with gender, we compare the race demographics per county with the Census data (Figure 2b). The correlation here is much stronger (0.56); however, the Twitter estimate greatly underestimates the minority population in majority-minority counties.

6.3 Reweighting by Race and Gender

Finally, we reweight our data based on these inferred demographics, using standard survey weighting (Gelman, 2007). If p is the true proportion of a category in a county, and $\hat{p}$ is our estimated proportion from Twitter, then the weight is simply $\frac{p}{\hat{p}}$. We compute weights for Female and Afro-Hispanic. For each user in our data for which we can infer gender and/or race, we adjust their contribution to the final feature vectors for that county according to these weights. For example, if a county is 60% female, but our Twitter estimate is 30% female, then tweets from each female in this county will effectively be counted twice. Otherwise, we use the same regression model described earlier to predict the health statistics of each county.

Figure 4 reports the relative reduction in SMAPE obtained by reweighting. We see that

for 20 of the 27 variables, reweighting results in more accurate models. On average, the relative reduction in SMAPE is 4.3%.

7. Discussion

These results suggest that adjusting for selection bias can greatly improve the accuracy of estimates made using social media data. These results hold despite the noise introduced by demographic inference.

As social media data become used more frequently, it will become increasingly important for researchers to at least measure, if not adjust for such bias. While it is tempting to consider a "system-wide" adjustment based on the overall Twitter population, most studies perform some additional selection — e.g., based on keyword, location, or time. Thus, other studies may exhibit even greater bias than ours, making such adjustments even more important.

Finally, we have used very simple demographic inference techniques here. In future work, we will consider more sophisticated demographic inference techniques applied to a wider range of attributes (Rao et al., 2011; Al Zamal et al., 2012; Mohammady and Culotta, 2014).

References

F Al Zamal, W Liu, and D Ruths. Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. In *ICWSM*, 2012.

M. Keith Chen. The effect of language on economic behavior: Evidence from savings rates, health behaviors, and retirement assets. *American Economic Review*, 103(2):690–731, April 2013. ISSN 0002-8282. doi: 10.1257/aer.103.2.690. URL https://www.aeaweb.org/articles.php?doi=10.1257/aer.103.2.690.

Barry R Chiswick and Paul W Miller. *The economics of language international analyses*. Routledge, London; New York, 2007. ISBN 9780203963159 0203963156. URL http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=186724.

Aron Culotta. 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*, 2010.

Aron Culotta. Lightweight methods to estimate influenza rates and alcohol sales volume from twitter messages. *Lang. Resour. Eval.*, 47(1):217238, March 2013. ISSN 1574-020X. doi: 10.1007/s10579-012-9185-0. URL http://dx.doi.org/10.1007/s10579-012-9185-0.

Aron Culotta. Estimating county health statistics with twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2014.

D D Danner, D A Snowdon, and W V Friesen. Positive emotions in early life and longevity: findings from the nun study. *Journal of personality and social psychology*, 80(5):804–813, May 2001. ISSN 0022-3514. PMID: 11374751.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. In *ICWSM*, 2013.

M. Dredze. How social media will change public health. *IEEE Intelligent Systems*, 27(4): 81–84, 2012. ISSN 1541-1672. doi: 10.1109/MIS.2012.76.

Maeve Duggan and Joanna Brenner. The demographics of social media users – 2012. Pew Internet & American Life Project, Feb 2013a.

Maeve Duggan and Joanna Brenner. *The demographics of social media users, 2012*, volume 14. Pew Research Center's Internet & American Life Project Washington, DC, 2013b.

Benito E Flores. A pragmatic view of accuracy measurement in forecasting. *Omega*, 14(2):93–98, 1986. URL http://ideas.repec.org/a/eee/jomega/v14y1986i2p93-98.html.

Daniel Gayo-Avello. Don't turn social media into another 'Literary digest' poll. *Commun. ACM*, 54(10):121128, October 2011. ISSN 0001-0782. doi: 10.1145/2001269.2001297. URL http://doi.acm.org/10.1145/2001269.2001297.

Andrew Gelman. Struggles with survey weighting and regression modeling. *Statistical Science*, 22(2):153–164, 2007.

Debarchana (Debs) Ghosh and Rajarshi Guha. What are we tweeting about obesity? Mapping tweets with topic modeling and geographic information system. *Cartography and Geographic Information Science*, 40(2):90–102, 2013. doi: 10.1080/15230406.2013.776210.

Louis A. Gottschalk and Goldine C. Gleser. *The Measurement of Psychological States Through the Content Analysis of Verbal Behavior*. University of California Press, January 1979. ISBN 9780520038134.

LE Graham, L Scherwitz, and R Brand. Self-reference and coronary heart disease incidence in the western collaborative group study. *Psychosomatic medicine*, 51(2):137–144, April 1989. ISSN 0033-3174. PMID: 2710908.

Carl L Hanson, Scott H Burton, Christophe Giraud-Carrier, Josh H West, Michael D Barnes, and Bret Hansen. Tweaking and tweeting: Exploring twitter for nonmedical use of a psychostimulant drug (adderall) among college students. *Journal of Medical Internet Research*, 15(4):e62, April 2013. ISSN 1438-8871. doi: 10.2196/jmir.2503. URL http://www.jmir.org/2013/4/e62/.

N. Heaivilin, B. Gerbert, J.E. Page, and J.L. Gibbs. Public health surveillance of dental pain via twitter. *Journal of Dental Research*, 90(9):1047–1051, September 2011. ISSN 0022-0345. doi: 10.1177/0022034511415273. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3169887/. PMID: 21768306 PMCID: PMC3169887.

Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H. Chi. Tweets from Justin Bieber's heart: The dynamics of the location field in user profiles. In *CHI*, pages 237–246, New York, NY, USA, 2011.

Ryan T. Howell, Margaret L. Kern, and Sonja Lyubomirsky. Health benefits: Meta-analytically determining the impact of well-being on objective health outcomes. *Health Psychology Review*, 1(1):83–136, 2007. ISSN 1743-7199. doi: 10.1080/17437190701492486. URL http://www.tandfonline.com/doi/abs/10.1080/17437190701492486.

Matthias R. Mehl James W Pennebaker. Psychological aspects of natural language. use: our words, our selves. *Annual review of psychology*, 54:547–77, 2003. ISSN 0066-4308. doi: 10.1146/annurev.psych.54.101601.145041.

Sue Jamison-Powell, Conor Linehan, Laura Daley, Andrew Garbett, and Shaun Lawson. "I can't get no sleep": Discussing #insomnia on Twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, page 15011510, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1015-4. doi: 10.1145/2207676.2208612. URL http://doi.acm.org/10.1145/2207676.2208612.

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

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

Vasileios Lampos, Tijl De Bie, and Nello Cristianini. Flu detector: tracking epidemics on twitter. In *ECML/PKDD*, pages 599–602, 2010. ISBN 3-642-15938-9, 978-3-642-15938-1. URL http://dl.acm.org/citation.cfm?id=1889788.1889832.

Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, , and J. Niels Rosenquist. Understanding the demographics of twitter users. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM'11)*, Barcelona, Spain, 2011.

Ehsan Mohammady and Aron Culotta. Using county demographics to infer attributes of twitter users. In *ACL Joint Workshop on Social Dynamics and Personal Attributes in Social Media*, 2014.

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

Michael J. Paul and Mark Dredze. You are what you tweet: Analyzing twitter for public health. In *Fifth International AAAI Conference on Weblogs and Social Media*, May 2011a. URL http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2880.

Michael J. Paul and Mark Dredze. You are what you tweet: Analyzing Twitter for public health. In *ICWSM*, 2011b. URL http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2880.

F. Pedregosa et al. Scikit-learn: Machine learning in python. *Machine Learning Research*, 12:28252830, 2011. URL http://scikit-learn.org/.

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

Lin Qiu, Han Lin, Jonathan Ramsay, and Fang Yang. You are what you tweet: Personality expression and perception on twitter. *Journal of Research in Personality*, 46(6):710–718, December 2012. ISSN 0092-6566. doi: 10.1016/j.jrp.2012.

08.008. URL http://www.sciencedirect.com/science/article/pii/S009265661200133X.

Delip Rao, Michael J. Paul, Clayton Fink, David Yarowsky, Timothy Oates, and Glen Coppersmith. Hierarchical bayesian models for latent attribute detection in social media. In *ICWSM*, 2011.

Adam Sadilek, Henry Kautz, and Vincent Silenzio. Predicting disease transmission from geo-tagged micro-blog data. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, December 2012. URL http://www.aaai.org/ocs/index.php/AAAI/AAAI12/paper/view/4844.

Matthias Schonlau, Arthur van Soest, Arie Kapteyn, and Mick Couper. Selection bias in web surveys and the use of propensity scores. *Sociological Methods & Research*, 37(3):291–318, February 2009. ISSN 0049-1241, 1552-8294. doi: 10.1177/0049124108327128. URL http://smr.sagepub.com/content/37/3/291.

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

Martin E. P Seligman. *Flourish: a visionary new understanding of happiness and well-being*. Free Press, New York, 2011. ISBN 9781439190753 1439190755 9781439190760 1439190763.

Alessio Signorini, Alberto Maria Segre, and Philip M. Polgreen. 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, May 2011. doi: 10.1371/journal.pone.0019467. URL http://dx.doi.org/10.1371/journal.pone.0019467.

Avar Stewart and Ernesto Diaz. Epidemic intelligence: For the crowd, by the crowd. In Marco Brambilla, Takehiro Tokuda, and Robert Tolksdorf, editors, *Web Engineering*, number 7387 in Lecture Notes in Computer Science, pages 504–505. Springer Berlin Heidelberg, January 2012. ISBN 978-3-642-31752-1, 978-3-642-31753-8. URL http://link.springer.com/chapter/10.1007/978-3-642-31753-8_55.