PREDICTING AND CHARACTERIZING THE HEALTH OF INDIVIDUALS AND COMMUNITIES THROUGH LANGUAGE ANALYSIS OF SOCIAL MEDIA

Johannes C. Eichstaedt

A DISSERTATION

in

Psychology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2017

Supervisor of Dissertation

____

Martin E.P. Seligman, Ph.D
Zellerbach Family Professor of Psychology

Graduate Group Chairperson

____

Sara Jaffee, Ph.D
Professor, Psychology

Dissertation Committee

Lyle H. Ungar, PhD
Professor of Computer and Information Science

Joe W. Kable, PhD
Baird Term Associate Professor of Psychology

To my Lieblingsfamilie, with all my love.

(Looks like I turned this into a Doktorarbeit.)

ACKNOWLEDGMENT

I thank Marty Seligman for his unwavering support, strategic advice and guidance—both explicit and implicit—across the last 6 years, during which we saw a small research collaboration with Google grow into the World Well-Being Project (WWBP), receive substantial funding, and become an institution. I thank Lyle Ungar foremost for being my friend/colleague, and probably teaching me most things I learned during the Ph.D. Thank you for letting me copy your cognitive structures, quick-witted cross-pollinating academic style, and good taste in science, both on my behalf and on that of my future students. I thank Andy Schwartz for the many things I've learned from him about how language works and how to think strategically about building not just code but an enduring computational infrastructure, but more so, for his tireless dedication to the cause of WWBP to measure the well-being and health of the world in space and time. His bravery of pinning down a snake in the Sierras and going hiking with me in the snow with temperatures in the teens were noted with astonishment and are fondly remembered.

I thank my parents for their love and unbroken belief in me despite what turned out to be 25 years of school across three countries, disciplines and graduate degrees—that would test the patience of any parent. I particularly thank them for their constant German encouragement to take more time off (which I did), and raising my amazing sister Christina who is as kind as she is supportive as she is smart, and always a source of deep perspective and travel finesse. On that note, I'd like to thank American Airlines for begrudgingly giving me status from the countless trips around the world over the last 5 years, and the National Park Service for their enduring commitment to the maintenance of the great long-distance trails across the U.S., and for always coming through on same-day permits whenever it was time to squeeze in a four-week hike.

I thank all 40+ members of WWBP who have served its overarching mission, with particular gratitude for Peggy Kern, who has shown us all how to be the world's most dependable colleague and collaborator; Greg Park, who set the bar at the very tippy top with his masterful data visualizations [require(great-greg-plots)] and tastefully composed manuscripts; Anneke Buffone, for her tireless dedication to pushing forward all 47 strategic objectives of WWBP at once, and for her bottomless generalized trust/empathy and non-specific prosociality; and Laura Smith for her loyal friendship and tireless dedication to WWBP, shouldering an unfathomable work load in helping boot up WWBP from an academic seedling into an institution.

I thank all my friends and the larger MAPP community for the perceived social support, but specifically David Yaden, for his original and surprisingly wise perspective, the untold and ineffable hilarity, and the patience shown across years of enduring my high-pitched mocking—you are the oldest young person I know; Alejandro Adler for his wide open heart that has and will improve the lives of millions of children around the world (Paula, thank you for marrying Ale, we all lucked out with that). I'd like to thank Zack Prager for being the coolest pro-musician pro-poker friend I have, for the many hours of Ram Dass discussions and for tagging along to the mountains and the lake (Megan, thank you again for marrying Zack. Good luck trying to make him fold clothes.); Roksana Filipowska, for being such a dear friend, firm believer, wise seer and counsel; Emlen Metz for her patient (if at times exasperated) friendship, the many years of acountabillibuddying and being the sort of friend you can always turn to when you need your premises checked; Peter Tremayne for showing us all how to hike the great trails of the world in your 70s and for being such a tireless trail angel in the Sierras, on behalf of all us backpackers; and Dan Buettner for his tireless curiosity, encouragement and overall high-impact awesomeness.

I'd like to express my gratitude for James Pawelski, proximally for the many quality walking breaks around the PPC during late hours and long weekends—my friend, you work too much—and distally for his indefatigable moral and philosophical leadership of the MAPP program and community. For Leona Brandwene, for teaching me some of her magic diplomacy skills, and Dan Bowling for good cheer and imperturbable perspective. I thank Kunal Sood for bridging the best the East and the West Coast have to offer, and opening some of the biggest stages of the world up to us faithful servants of Science.

I'd like to thank the Penn Psychology faculty for their excellent instruction, and especially Melissa Hunt, Dianne Chambless and Alan Goldstein for their kind, patient and in-depth instruction and supervision in the clinical program—what a privilege to have such capable and caring instructors. Thank you so much for investing your time in my training.

I'd like to thank Peter Schulman for his support, skillful administrative advice and perspective over the years, and the Positive Psychology Center for the excellent resources and infrastructure, and for letting me do my Ph.D. on a 4-screen mega-desk. I gratefully acknowledge the help of Courtney Hagan in formatting this dissertation; Joe Kable for the many good comments on my work and for serving on my committee, and Yuni Thornton for her level admin foo and wise counsel.

Finally, I'd like to thank my parents, the Templeton Religion Trust, the Trustees of the University of Pennsylvania and the German Academic Scholarship Foundation for the financial support of many $100,000s that went into rewiring the neurons of this one brain to make it better at Science. I shall try to do something good with it.

ABSTRACT

PREDICTING AND CHARACTERIZING THE HEALTH OF INDIVIDUALS AND COMMUNITIES THROUGH LANGUAGE ANALYSIS OF SOCIAL MEDIA

Johannes C. Eichstaedt

Martin E. P. Seligman

A large and growing fraction of the global population uses social media, through which users share their thoughts, feelings, and behaviors, predominantly through text. To quantify the expression of psychological constructs in language, psychology has evolved a set of "closed-vocabulary" methods using pre-determined dictionaries. Advances in natural language processing have made possible the development of "open-vocabulary" methods to analyze text in data-driven ways, and machine learning algorithms have substantially improved prediction performances. The first chapter introduces these methods, comparing traditional methods of text analysis with newer methods from natural language processing in terms of their relative ability to predict and elucidate the language correlates of age, gender and the personality of Facebook users (N = 65,896). The second and third chapters discuss the use of social media to predict depression in individuals (the most prevalent mental illness). The second chapter reviews the literature on detection of depression through social media and concludes that no study to date has yet demonstrated the efficacy of this approach to screen for clinician-reported depression. In the third chapter, Facebook data was collected and connected to patients' medical records (N = 683), and prediction models based on Facebook data were able to forecast the occurrence of depression with fair accuracy–about as well as self-report screening

surveys. The fourth chapter applies both sets of methods to geotagged Tweets to predict county-level mortality rates of atherosclerotic heart disease mortality (the leading cause of death in the U.S.) across 1,347 counties, capturing 88% of the U.S. population. In this study, a Twitter model outperformed a model combining ten other leading demographic, socioeconomic and health risk factors. Across both depression and heart disease, associated language profiles identified fine-grained psychological determinants (e.g., loneliness emerged as a risk factor for depression, and optimism showed a protective association with heart disease). In sum, these studies demonstrate that large-scale text analysis is a valuable tool for psychology with implications for public health, as it allows for the unobtrusive and cost-effective monitoring of disease risk and psychological states of individuals and large populations.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER FOUR

PREFACE

All of the work presented in this dissertation was conducted at the World Well-Being Project (WWBP) at the Positive Psychology Center at the University of Pennsylvania. All studies were approved by the University of Pennsylvania Institutional Review Board. The analyses of chapters 1, 2 and 4 are based on the WWBP Python code base, a large part of which has been released open-source [Schwartz, H. A., Giorgi, S., Sap, M., Crutchley, P., Eichstaedt, J. C., and Ungar, L. H. (2016). Differential Language Analysis Toolkit 1.0.] (see dlatk.wwbp.org).

An earlier version of Chapter 1 was written as a review for one of my qualifying examinations at the end of the third year. I have continued to serve as the lead investigator responsible for all concept formation, data analysis, as well as manuscript composition. M. L. Kern, D. B. Yaden, V. Tobolsky, C. A. Hagan and J. Iwry have contributed to manuscript edits. H. A. Schwartz, G. Park and L. H. Ungar gave feedback about manuscript scope and focus.

Chapter 2 is an invited submission to *Current Opinion in Behavioral Science* which I was invited to submit as its senior author. I served as the lead investigator, responsible for review structure and organization and the majority of manuscript composition. S. C. Guntuku and D. B. Yaden wrote the manuscript with me, M. L. Kern and L. H. Ungar provided revisions.

I was the lead investigator for the project discussed in Chapter 3, but not responsible for data collection in the Emergency Department (discussed in Padrez et al., 2015). I was responsible for all major areas of concept formation, analysis and result composition and the majority of manuscript composition. H. A. Schwartz and P.

Crutchley created the majority of the computational infrastructure used in the methods. R. J. Smith, V. A. Tobolsky and H. A. Schwartz contributed to manuscript composition. D. Preoţiuc-Pietro, M. L. Kern, L. H. Ungar and R. M. Merchant provided revisions.

An earlier version of Chapter 4 was written as the culmination of my 699 first year research project and has been published [Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., Jha, S., Agrawal, M., Dziurzynski, L. A., Sap, M., Weeg, C., Larson, E. E., Ungar, L. H., & Seligman, M. E. (2015). Psychological Language on Twitter Predicts County-Level Heart Disease Mortality. *Psychological Science*. *26(2)*, 159-169.]. The version given here is the last version before copy edits, reprinted by permission of SAGE publications. I was the lead investigator and led the project; I and H.A. Schwartz conceived of the study; H. A. Schwartz, I, G. Park, S. Jha, M. Agrawal, L. A. Dziurzynski, and M. Sap handled data acquisition, processing, prediction model development, and data analyses; I, M. L. Kern, H. A. Schwartz, and G. Park drafted the manuscript; D. R. Labarthe, R. M. Merchant, L. H. Ungar, and M. E. P. Seligman provided critical revisions. C. Weeg and E. E. Larson helped acquire process and analyze county-level information.

The Three Theorems of Psychohistorical Quantitivity:

The population under scrutiny is oblivious to the existence of the science of Psychohistory.

The time periods dealt with are in the region of 3 generations.

The population must be in the billions (±75 billions) for a statistical probability to have a psychohistorical validity.

—Isaac Asimov, Foundation, 1966

INTRODUCTION

Over the last two decades, “those of us who use computers, and other networked devices have become a part of an emerging longitudinal, cross-sectional, and cross-cultural study” (Illiev, Dehghani, & Sagi, 2014, p. 21). Specifically, the digitization of social life, in the form of social media, has resulted in a massive repository of natural language associated with specific individuals. Much of this data is public (Twitter), and that which is private can often be accessed at large scale through electronically distributed consent forms and collection systems (such as Facebook applications).

In *Clinical vs. Statistical Prediction: A Theoretical Analysis and Review of the Evidence*, Meehl (1954) changed psychology by demonstrating the superiority of “mechanical” or statistical modes of prediction over subjective, intuitive judgments. Since the publication of Meehl’s article, self-report scales have become the de-facto standard for psychological assessments, and standards have emerged regarding reliability, validity, factor analytic, and other psychometric properties. This dissertation describes a mechanical mode of prediction that substantially extends psychometric self-report methods to unobtrusively assess large fractions of populations.

The capacity for and habit of communicating through language is a fundamental component of human behavior. Psychology has a long history of using automated language analysis to try to measure psychological states using pre-determined and often theory-based dictionaries. In *The Secret Life of Pronouns*, Pennebaker (2011) shows how such traits as gender and personality can be predicted through syntactic "filler" words which are difficult to detect or control in speech or writing, suggesting that how we use language encodes underlying psychological processes. Advances in Natural Language Processing (NLP) in computer science now allow algorithms to generate highly interpretable yet theoretically agnostic data-driven language variables that can be used to analyze language with large conceptual and behavioral resolution. In conjunction with advances in machine learning—the modern set of statistical tools that has enabled voice-operated assistants on our smartphones and self-driving cars—these types of computational language analyses, when applied to social media datasets, have effectively provided psychology with mechanical modes of prediction that extend, and in some cases step well beyond, self-report measures (Kosinski, 2014; Kosinski, Stillwell, & Graepel, 2013).

In order to introduce and demonstrate the predictive power of these methods, I begin this dissertation with an overview of old and new methods of language analysis (chapter 1). I then apply language analysis and machine learning to Twitter and Facebook data sets to predict and characterize the most prevalent physical illness and mental disorder: In chapter 2, I predict and characterize the psychological determinants of heart disease rates of communities; in chapters 3 and 4, I discuss the use of social media to predict the depression status of individuals. Across the following chapters, I demonstrate that large-scale text analysis is a valuable tool for psychology and allows for the unobtrusive, cost-effective, non-reactive monitoring of psychological states for both individuals and large populations.

Social Media

Psychologists have long turned to "behavioral residues" (Gosling, Ko, Mannarelli, & Morris, 2002) to understand the psychological states of individuals. With the digital revolution, data sets have become available that encompass large portions of populations, rather than narrow study samples. As of 2017, Google's email service Gmail has 1 billion (Gibbs, 2016) and Facebook has 1.86 billion monthly active users (Facebook: Our Mission, n.d.). Among these *big* data sources, social media stands out as a source of autobiographical text that has disclosure of thoughts, emotions and behaviors as its goal (Kramer, 2010). Social media data is public by design (like Twitter), or accessible to researchers through targeted data collection through apps (like Facebook; e.g., Kosinski & Stillwell, 2012). Other big data sources (like search queries) can certainly be mined to detect individual-level markers of psychological states and illness (e.g., Yom-Tov, White, & Horvitz, 2014) and population trends in physical (e.g., the flu; Butler, 2013) and mental health (e.g., depression; Yang, Huang, Peng, & Tsai, 2010). However, while definitive empirical comparisons of the value of different large-scale data sources are still missing from the literature, nothing seems to compare to the richness of self-disclosure observed on social media and publication trends in psychology seem to confirm this view (see Figure 1).

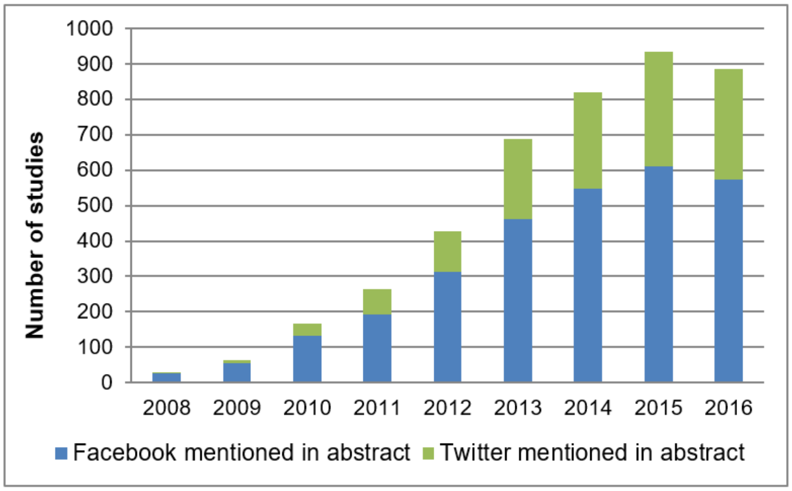


Figure 1. Number of studies indexed by PsycINFO mentioning Facebook (blue) or Twitter (green) in their abstract between 2008 and 2016 (as of March 2017, 2016 indexing not complete).

Text Analysis

The beauty of text as a variable is that it is intrinsically and immediately interpretable. In technology parlance, recording human thought as text is a “proven technology,” going back at least to the Cuneiform script on clay tablets invented by the Sumerians in the 3rd millennium BC (Zimerle, 2010). In principle, given a sufficient number of clay tablets and outcome data (e.g., harvest records), the open-vocabulary methods discussed in this dissertation (specifically, Differential Language Analysis) could be used to characterize the cultural goings-on of good harvest years in ancient Sumer. As such, these methods are fundamentally applicable to all written language, perhaps the defining cultural practice of our species.

However, social media sites cover a number of different “feature sets” beyond the text content of users’ posts, which range from activity meta data (when is content posted) to data that captures the platform-specific social graph (who is Facebook friends with whom, who retweets whom on Twitter) to the content of images and other more platform-specific features (such as Facebook likes). All of these feature sets have been shown to contain relevant information to predict psychological states or traits of users (e.g, meta-features and social graph on Twitter: De Choudhury, Gamon, Counts, & Horvitz, 2013, Facebook likes: Kosinski, Stillwell, & Graepel, 2013, images: Liu, Preotiuc-Pietro, Samani, Moghaddam, & Ungar, 2016). Interpretations based on these feature sets, however, seem to have limited generalizability beyond the context of the platform in question, and thus limit the external validity of studies that critically rely on them. Do reciprocal retweets really mean that two users are "friends"? Or that re-sharing other users' links is a sign of "social engagement?" Social media platforms will come and go with every generation as will likes and retweets, yet text is here to stay.

The Usefulness of Prediction Models

Many of the defining papers in the young field of social-media-based big-data psychology present as their central contribution (Kosinski, Stillwell, & Graepel, 2013, Park et al., 2015) or incorporate (Schwartz et al., 2013b) performances of machine learning models predicting psychological characteristics from social media data. To psychologists, who are primarily motivated by obtaining psychological insight, it may not be immediately obvious why prediction accuracies matter.

I argue that prediction performances ought to be best understood as gauges of how much variance of a psychological construct is captured in a given feature set (in our case, predominantly text) in the context of how much variance is accounted for by other predictors (such as demographics). In many of the publications of our research group, language use is analyzed for psychological insight, as a data-driven method to characterize the emotional, cognitive and behavioral correlates of a particular psychological construct. For this kind of analysis, prediction performances are an important complement to help us understand how seriously we should take the particular language markers used for psychological insight. If a language-based prediction model does not add predictive performance beyond a model using demographics or income, we ought to assume that most of the language markers observed are related to demographics or socioeconomic status. In models where language-based predictions add additional variance to gold standard models that combine demographics, socioeconomics and health risk factors, we may be hopeful that the language markers will tell us something about psychological characteristics over and above these other factors. Of course, various methods of statistical control can and should be used to adjust for the covariance of specific language features with these other variables, but a comparison of overall prediction performances gives an important first estimate on how much one might to expect to learn from the language-based analyses.

This Dissertation

Chapter 1: Open and Closed-Vocabulary Methods in Computational Linguistic Analysis. In the first chapter, I review methods of computerized language analysis in psychology. Text analysis for psychological insight has traditionally relied on theory-driven “closed-vocabulary” analysis programs, which restrict analysis to words from predetermined dictionaries. Methods from Natural Language Processing offer data-driven "open vocabulary" discovery and classification of psychological constructs in text. I then provide a direct comparison of the three most popular dictionary-based programs (the General Inquirer, DICTION and Linguistic Inquiry and Word Count [LIWC] 2015) and two open-vocabulary methods (topic modeling and Differential Language Analysis of words and phrases). I apply these approaches to the Facebook statuses of N = 65,896 Facebook users who have taken a Big Five personality inventory to compare the respective language correlates of user age, gender, and personality traits across methods. I find substantial overlap between the dictionary-based programs in the concepts covered by their dictionaries, but also that highly frequent and ambiguous words may dominate dictionary associations. Open-vocabulary methods help to specify and disambiguate dictionary findings and prevent such mistakes in the analysis, while offering finer and more transparent units of analysis. Using language variables in regression models, I find that LDA topics capture significantly more outcome-related variance than the closed-vocabulary approaches. I conclude that dictionary-based programs continue to offer valuable information to psychologists interested in text-analysis, especially with regard to pronoun use and other function words as indicators of underlying attentional and emotional processes. However, more specific and transparent units of analysis of open-vocabulary approaches are preferable in data sets with thousands of observations for data-driven exploration of language correlates. I conclude by providing guidelines for choosing linguistic analysis methods.

Chapter 2: Detecting Mental Illness Through Social Media: A Review. In the following two chapters, I discuss the use of social media to detect (i.e., predict) the mental health status of individuals. The second chapter provides a review of the existing literature, across Facebook, Twitter and web forums as a source of text. In these studies, mentally ill users are identified using screening surveys, their public sharing of a diagnosis on Twitter, or by their membership in an online forum, and they are distinguished from control users by patterns in their language and online activity. Linguistic analysis methods may help to identify at-risk, depressed individuals through large-scale passive monitoring of social media. However, at this point there are no studies published that use assessments of the mental health status of the social media users based on something other than self-report. In the third chapter, I present the results from such a study, in which the depression status is determined by clinician judgment as recorded in medical records.

Chapter 3: Predicting Depression Through Facebook. This study examines the Facebook language correlates of depression in a real-world medical setting, as well as predict its occurrence in the medical record. 683 patients visiting a large, urban, academic emergency department consented to a collection of their history of Facebook statuses in conjunction with their medical records. Prediction models were trained on the language data collected preceding the first recorded diagnosis of depression of 114 depressed patients, and every depressed patient was matched with five patients without a diagnosis of depression, for whom Facebook data from the same time span was considered. Facebook-language-based models can predict the first recording of depression in the medical record with fair accuracy, and about as well as the accuracy of screening surveys reported in another study. Our results suggest that machine learning applied to social media language can both identify individuals at risk for depression and improve existing screening and monitoring procedures.

Chapter 4: Predicting Heart Disease through Twitter. While the first three studies discuss prediction of health status at the individual level, the study presented in the fourth chapter generalizes prediction through social media to the community level. In this chapter, I present a study that uses Twitter language to predict mortality of atherosclerotic heart disease (AHD) at the county level, and explore it its psychological correlates. Language patterns reflecting negative social relationships, disengagement, and negative emotions—especially anger—emerged as risk factors; positive emotions and psychological engagement emerged as protective factors. Most correlations remained significant after controlling for income and education. A cross-sectional regression model based only on Twitter language predicted AHD mortality significantly better than did a model that combined 10 common demographic, socioeconomic, and health risk factors, including smoking, diabetes, hypertension, and obesity. Capturing community psychological characteristics through social media is feasible, and these characteristics are strong markers of cardiovascular mortality at the community level.

CHAPTER 1

OPEN AND CLOSED-VOCABULARY METHODS IN COMPUTATIONAL LINGUISTIC ANALYSIS

Digital text has become the predominant form of human communication across the world. In the last decade, “those of us who use computers, and other networked devices have become a part of an emerging longitudinal, cross-sectional, and cross-cultural study” (Illiev, Dehghani, & Sagi, 2014, p. 21). This real-world study encompasses large fractions of populations, which moves far beyond the narrow study samples that have typified psychological studies for the past two centuries. In the age of information, massive datasets are constantly being generated. One such pool of data comes from the words written by users on social media, such as Twitter and Facebook. The mass public engagement with these platforms provides an unprecedented opportunity to study the psychological experience of millions of people.

Humans have a long history of creating written records of their thoughts, behaviors, and experiences, and psychology has a long history of analyzing such texts for psychological insight. Text analysis in psychology began with systematic content analysis: manualized coding systems instructed human raters how to assign codes to passages of text based on the occurrence of certain "themes", which were then translated into insights regarding the presence or absence of a stipulated psychological construct (Mehl, 2006). Early examples include the psychoanalytical coding of responses to the Rorschach Inkblot Test (Rorschach, 1942) and the Thematic Apperception Test (Morgan & Murray, 1935). Systematic approaches arose through the 1960s and 70s, with qualitative methodologies such as grounded theory (Glaser & Strauss, 1967) being developed. More recently, the Content Analysis of Verbatim Explanations (CAVE) coding system was developed to capture the authors' explanatory style (Peterson & Semmel, 1982; Peterson, Luborsky, & Seligman, 1983; cf. Smith, 1992 for an overview of this and 13 other coding systems).

With the availability and increasing bandwidth of computers, the possibility arose that the coding process could be expedited and human coder bias could be removed. Computerized text analysis was first introduced about fifty years ago, with various programs developed over successive decades. At their core, these programs reduce words to numbers. These programs employ theory-driven "dictionaries," or list of words assigned to a specific category, scanning a text, counting the occurrence of words within that category, and outputting the relative frequency (percentage) of words in the text contained in that dictionary. Among these programs, the General Inquirer (Stone, Dunphy, Smith, & Ogilvie, 1966), DICTION (Hart, 1984) and Linguistic Inquiry and Word Count (LIWC; Pennebaker, Booth, & Francis, 2007) have received the most attention in the literature.

These text analysis programs are straightforward and useful for simple quantification. Over the past two decades, methods borrowed from Natural Language Processing (NLP), such as Latent Semantic Analysis (LSA; Landauer & Dumais, 1997) and its more sophisticated successor Latent Dirichlet Allocation (LDA; Blei, Ng, & Jordan, 2003), have been introduced to psychological research. Rather than relying on existing dictionaries, these newer methods allow for the data-driven discovery of patterns in text. Despite excellent reviews introducing such approaches to psychological audiences (c.f. Griffiths, Steyvers, & Tenenbaum, 2007; Landauer & Dumais, 1997), these methods require substantially more technical and statistical sophistication than the traditional text analysis programs, and have only recently started to be applied more in the psychological literature.

The different closed-vocabulary dictionaries and growing number of open-vocabulary approaches provide different tools that might be useful at different times, depending on one's purpose. This review aims to provide empirical guidance as to which tool is most appropriate for different circumstances. We first introduce closed and open vocabulary methods. Then, we quantitatively compare the performance of traditional text analysis programs and the data-driven methods from NLP on a large dataset of Facebook status updates. We conclude by providing guidelines for choosing linguistic analysis methods across different research contexts.

Closed-Vocabulary Method

The simplest way to describe language use quantitatively is to count the number of times individual words occur relative to the total number of words in a text. For example, "I walked outside and I enjoyed the warm sunshine" contains nine words, giving *"sunshine* a relative frequency of 11.1%, and *I* a relative frequency of 22.2%. Related words can be combined in *dictionaries*, researcher-created lists of words that are theoretically presumed to have something in common, like indicating positive emotion or being personal pronouns. A verb dictionary might include 500 words, such as *walked* and *enjoyed*, and a "verb score" can be calculated by summing the relative frequencies of the verbs contained in the dictionary (22.%). Once these dictionary-based relative frequencies are derived for different texts, they can be compared to one another and correlated with other variables using the usual methods of inferential statistics. For example, women are more likely to use social words than men (Newman, et al., 2008). The dictionary-based word-count approach is a seemingly transparent way to generate statistically meaningful language variables and is used by all major text analysis programs in psychology (Mehl, 2006).

Closed-vocabulary text analysis programs. Based on previous reviews (e.g., Neuendorf, 2002), we compiled a list of 31 text analysis programs.[1] Of these, only six are designed to track specific psychological dimensions based on included dictionaries (rather than provide a generic infrastructure for counting keywords) and have more than a few citations in the academic literature: the General Inquirer (GI; Stone et al., 1966), DICTION (Hart, 1984), Regressive Imagery Dictionary / Count (Martindale 1973, 1975), TAS/C (Mergenthaler & Bucci, 1999), Gottschalk-Gleser Scales / Psychiatric Content Analysis and Diagnosis (PCAD 2000; Gottschalk & Bechtel, 1995), and Linguistic Inquiry and Word Count (LIWC; Pennebaker et al., 2007).

The differences between different programs predominantly concern the number and quality of the included dictionaries. Of these six programs, three (GI, DICTION, and LIWC) are designed to carry out text analysis across a large number of dimensions, and thus we review these programs in greater detail in historical order. The other three are designed for narrower application in clinical or psychoanalytic contexts and are omitted from further discussion. Of the three included programs, LIWC has had by far the largest

[1] ACTORS, CATPAC, CONCORD, Concordance 3.3, Count, CPTA, Diction 7.0, DIMAP-4, General Inquirer, Hamlet, IDENT, Intext 4.1 (now TextQuest 4.2), Lexa, LIWC, MCCALite, MECA, MonoConc, ParaConc, PCAD 2000, PROTAN, SALT, SWIFT, TABARI, TAS/C, TextAnalyst, TEXTPACK, TextSmart, The Yoshikoder, VBPro, WordStat 6.1.

impact in the literature in Google Scholar as of March 2017, with 4,500 citations (for Pennebaker, Francis, & Booth, 2001; Pennebaker, 1997a; Pennebaker, 1997b), followed by the General Inquirer with 2,100citations (Stone, Dunphy, Smith, & Ogilvie, 1966; Kelly & Stone, 1975; Stone, Bales, Namenwirth, & Ogilvie, 1962), and Diction with 600 (Hart, 1984; Hart, 2001; Hart, 1997).

The General Inquirer. The General Inquirer (GI) was developed at Harvard University in the 1960s for mainframe computers and was used most frequently during the 1960s and 70s. As the program was designed during the early days of computing, tape drives provided memory and key cards were used to input data into a mainframe environment. It was designed for general, multi-purpose text analysis, but could also extract custom dictionaries (Stone, Bales, Namenwirth, & Ogilvie, 1962). Over 25 dictionaries were designed between 1962 and 1965. Users were cautioned against having "unrealistic expectations" about the ease of use (Kelly & Stone, 1975, p. 112), yet the program set the standard for computerized programs that followed.

The latest version of the General Inquirer (http://www.wjh.harvard.edu/~inquirer/), includes 182 dictionaries (see Online Supplement 1), split into three main sets: 63 Lasswell Dictionaries, 107 Harvard Psychosociological Dictionaries, which include seven dictionaries intended to help with word sense disambiguation and five social cognition dictionaries distinguishing different verb and adjective types, and 12 Stanford Political Dictionaries (the same word can appear in multiple dictionaries). Considerable resources were invested in the construction of the GI dictionaries, with more than 10,000 human rated annotations collected for the 12 Stanford Political Dictionaries alone (Stone et al., 1966).

Lasswell dictionaries. A first set of dictionaries were designed to measure eight value domains stipulated by Lasswell and Kaplan's (1950) influential book, *Power and Society: A Framework for Political Inquiry*, and included four deference categories (*power*, *rectitude, respect, affection*) and four welfare categories (*wealth, well-being, enlightenment, skill*; Lasswell & Namenwirth, 1969). Each of these eight categories were subdivided into three dictionaries: *participants*, *transactions* (i.e., social allocation, or processes pertaining to the social distribution of values), and *other* words, as well as a *total* dictionary that contains all words across *participants, transactions,* and *other* in a given domain (cf. Weber, 1984, 1990). For example, under the category of *wealth*, the *participants* dictionary included *company*, *bank*, and *customer*; the *transactions* dictionary included *spend, bought*, and *raise*, and the *other* dictionary included *car, own*, and *money*. Additional dictionaries were later added to cover other processes not covered by Lasswell's theory.

Harvard psychosociological dictionaries. A second set of dictionaries were designed as a general set of dictionaries that could extract information relevant to the leading psychological (e.g., Morgan & Murray, 1935; Murray, 1938, 1943) and sociological theories (e.g., McClelland, 1961) of the day. For example, McClelland, Davis, Wanner and Kalin (1966) used these dictionaries to study the connection between folklore and drinking in a sample of 44 primitive cultures. The dictionaries have undergone several updates, with the most recent form being the Harvard Psychosociological IV Dictionary (107 dictionaries).

Stanford political dictionaries. A third set of dictionaries were designed to explore the assertion that decision-making can be measured along three dimensions:

evaluation (positive--negative), potency (strong--weak) and activity (active--passive), (Osgood, 1963; Osgood et al., 1957). Every word was assigned to and weighted along one, two, or three of these dimensions (e.g., calm is positive affect + weak + passive) by multiple human judges. The Stanford dictionaries covered 98% of the words encountered in texts of the time (Stone et al., 1966). For example, Holsti, Brody, and North (1964) used these dictionaries to analyze the available verbatim text recorded from the key decision makers during the Cuban missile. During the most heated part of the conflict, “strong-active-negative” perceptions of the adversary prevailed on both sides. As the conflict was resolved, the American perception first became more neutral (more “positive” and less “negative”) during the bargaining period (beginning October 25th), and then the Russian perceptions of the Americans followed suit on October 27th.

DICTION. DICTION was developed in the 1980s to analyze the “verbal tone” in 500 word selections from US presidential speeches (Hart, 1984). DICTION assumed that political texts could be characterized according to five master variables -- *activity, certainty, commonality, optimism,* and *realism* – such that “if only five questions could be asked of a given passage, these five would provide the most robust understanding” (Hart, 2001, p. 45). Each master variable was then composed of adding and subtracting the frequencies of multiple dictionaries.

In its current form, DICTION employs 31 non-overlapping dictionaries, containing 9,334 terms, as well as four variables (*Complexity*, *Embellishment, Insistence, Variety*) that encode relative lengths of words, ratio of adjectives to verbs, relative frequency of words repeated more than three times out of every 500 words, and the ratio of unique to total words, respectively. These 35 language variables are then combined into the five “master” variables by adding and subtracting their standardized (Z) scores from one another. For example, *Certainty* is derived by adding the standardized scores of *tenacity*, *leveling*, *collectives* and *insistence*, and by subtracting *numerical terms*, *ambivalence*, *self-reference* and *variety*. For all master variables, a constant of 50 is added to the result, to eliminate negative numbers. DICTION includes norm scores, which were developed from various texts, and the master variable scores of a given text can be compared to these baselines. DICTION also allows custom dictionaries.

Linguistic Inquiry and Word Count. The Linguistic Inquiry and Word Count (LIWC) program was originally designed in 1993 to analyze a collection of essays written during expressive writing interventions (Francis & Pennebaker, 1992, 1993; Pennebaker, Francis, & Booth, 2001; Pennebaker et al., 2007). The program has subsequently been applied to texts across a variety of domains and identified consistent patterns.

LIWC relies exclusively on word count and ignores word order and any factors other than relative frequency of dictionaries in a given text. The latest version (LIWC2015) was recently made available, and aims to allow a simple and easy to use flexible option for analyzing English and non-English word samples. LIWC is organized hierarchically, with some dictionaries subsuming others. General categories include function words, grammar, affect words, social words, cognitive processes, perpetual processes, biological processes, core drives and needs, time orientation, relativity, personal concerns, informal speech, and punctuation. These dictionaries are further split into multiple dictionaries. For instance, the *affective* dictionary is further broken into *positive emotion* and *negative emotion*, with *sadness*, *anxiety*, and *anger* sub-dictionaries.

As a result, when sub-dictionaries (like *sadness*) correlate with an outcome, this often drives a correlation between the outcome and a higher order dictionary (like *affective processes*). Output also provides summary variables, including word count, and metrics based on linear combinations of dictionary frequencies (like emotional tone).

LIWC's primary contribution rests in its distinction between "function" and "content" words (Chung & Pennebaker, 2007). Function words (often also referred to as "style" words) provide the syntactic scaffolding of language; they consist of pronouns (s*he, I, we*), articles (*the, an, a)*, prepositions (*of, as, by),* and conjunctions (*and, or, so)*. There are fewer than 200 common function words in the English language, yet they represent over half of all words used (Mehl, 2006). Content words include nouns (*book, stage, park*) and non-auxiliary verbs (*swimming, snowing, sleeping*). There are many more content words and dictionaries, but they are used less frequently. For instance, the set of words LIWC includes in its *emotional* dictionaries accounts for less than 5% of the language used in everyday writing, including poetry (Mehl, 2006). According to Mehl (2006), function words are indifferent to content and are typically used without conscious attention. Their high relative frequencies of occurrence make function words particularly suitable units of analysis. Part of the success of LIWC lies in its ability to find patterns in pronoun use (e.g., Campbell & Pennebaker, 2003; Chung and Pennebaker, 2007; Pennebaker, 2011).

Benefits and limitations of closed-vocabulary methods. The closed-vocabulary methods implemented by GI, DICTION, and LIWC is are a theory-driven, top-down approach: the text is scanned for the occurrence of specific words, which were previously assigned to dictionaries intended to measure various theoretical constructs. This approach is responsible for the majority of published findings on psychological correlates of language. The main advantage of this approach is that it transforms the thousands of mostly rarely used words in a given text sample into 10-100 interpretable language variables that can be explored with standard statistical techniques, and that the derived language variables are comparable across studies.

Despite their benefits and wide-spread use in the psychological literature, they also bring numerous challenges (see also Kern et al., 2016). Dictionaries such as these are rigidly defined and are not altered in response to the data to which they are applied; their vocabularies are "closed" and "theory-driven." They are insensitive to context, and reduce text to a statistical *bag of words*, which is indifferent to word order. Each word is matched against dictionaries individually. Negation is ignored, such that the phrase "I am not happy" is scored as 25% *positive emotion.* Further, this method cannot clarify lexical ambiguities (words appearing in different parts of speech and/or with different senses). For example, a *belt* may both be worn and be the home of asteroids. The open-vocabulary approaches described below alleviate some, but not all, of these limitations, as will be discussed below.

It is also worth considering a fundamental challenge of working with language. Whereas most psychological variables are assumed to be normal, the frequency distribution with which words are used is *extremely* skewed. Specifically, the relative frequency of words in a language follows Zipf's law (Pierce, 1980), which stipulates that the probability of encountering the rth most common word in a given language is inversely proportional to its rank (r) in that language for some normalization constant k:

$P(w_r) \sim \frac{k}{r}$. In other words, the frequency of the rth most frequent word is given by $P(w_r) = \frac{0.1}{r}$, until about rank 1,000, such that the most common word (in English: *the*) would have a probability of occurrence of $P(w_1)$ = .1 = 10%, followed by *to* with 5%, and so forth. The vast majority of words are in the long tail of the distribution and will only be used by a small fraction of a given sample. This accounts for Mehl's assertion (2006) that there are fewer than 200 common function words, yet they represent over half of all words used.

As an example, Figure 1 shows the frequency distribution of the 500 most frequent words in this sample from 65,896 Facebook users. Beyond the very common words that fulfill mostly syntactic roles (articles, pronouns, prepositions and conjunctions), most words occur very rarely. Even when limiting the sample to words that are used by at least 1% of the users in the sample, there remain 9,570 unique words across 258 million word instances. The most frequent 96 words account for more than 50% of word occurrences, and the top 1,000 words for more than 82% (See Figure 1b).

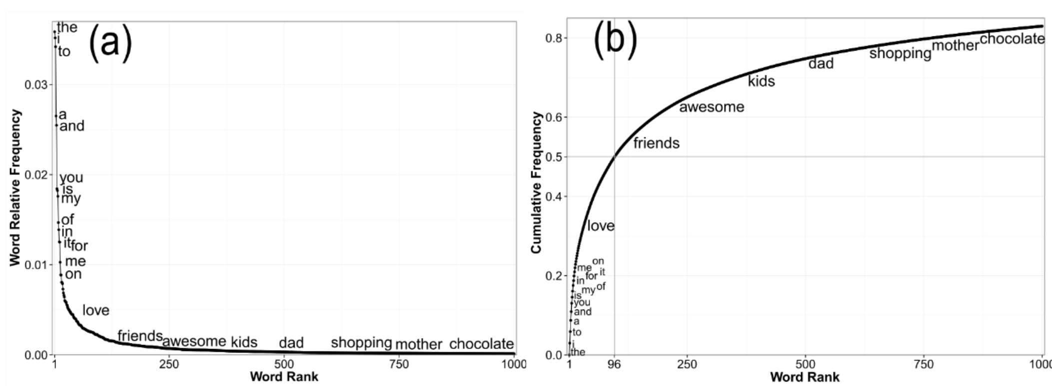


Figure 1. The relative frequency of the 1,000 most common words in a language sample from 65,896 Facebook users, shown (a) as a typical Zipfian distribution, in which the frequency of a word is inversely proportional to the word's frequency rank within a given

language, and (b) as the cumulative frequency of the most common 1,000 words in the sample.

Because of this distribution of words, single words make poor units of analysis unless very large language samples are available. The three closed-vocabulary methods described above try to get around this by grouping words together into meaningful categories. However, the distribution of word frequencies implies that one or two words can completely dominate the overall frequency of a particular dictionary, and thus the observed correlation of the dictionary with another variable. Further, the established dictionaries make no attempt at disambiguating different word senses, nor take their relative frequencies into account, which may shift over time. For example, LIWC includes the word “sick” in the *negative emotion* and *biological* dictionaries. And yet for many young people, “sick” is increasingly used to indicate that something very desirable. The closed-vocabulary dictionaries are insensitive to word sense ambiguities and such semantic drift.

Open-Vocabulary Methods

As an alternative to theory-driven dictionaries, various techniques from NLP can be used on language data to reduce the number of dimensions from thousands of words to a manageable set of factors, and do so with full transparency about which words drive which factors.

Among these data-driven “open-vocabulary” approaches, Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) have received the most attention in the psychological literature (cf. Schwartz & Ungar, 2015). A full review of LSA and LDA is beyond the scope of this article (for excellent reviews see Griffiths, Steyvers, &

Tenenbaum, 2007 and Landauer & Dumais, 1997;). Here, we briefly introduce the methods, and add a discussion of Differential Language Analysis (DLA), an exploratory technique developed and introduced to psychology by our group (e.g., Schwartz et al., 2013b), which is based on the use of LDA topic models and relative frequencies of words and phrases.

Latent Semantic Analysis (LSA). LSA was first developed in the late 1980s to determine the similarity between two bodies of text (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990; Dumais, Furnas, Landauer, Deerwester, & Harshman, 1988). LSA is similar in nature to factor analysis, which is frequently used in psychology to reduce a large number of independent variables (e.g., many survey items) to a smaller number of latent factors that account for a large fraction of the variance. A factor analysis might be applied to a matrix in which columns are items, the rows are different participants, and cells are the participants' responses to the items. A similar matrix can be created for language analysis, in which the columns index different language documents (e.g., transcripts, or as in the present study, Facebook statuses) and the rows index different words. A cell in this matrix would thus give the number of times a word is used in a given document. This word-by-document (WBD) matrix can then be factor-analyzed using singular value decomposition (SVD), yielding a set of latent semantic factors. (SVD is a factorization technique similar to Principal Component Analysis; see Landauer & Dumais, 2007 for a full review of LSA.)

Classical psychological factor analysis yields an approximation of the participant-by-item matrix that expresses (a) a participant's' responses as a combination of factor scores, and (b) survey items as loadings on factors. LSA yields an approximate representation of the WBD matrix that expresses (a) documents as combinations of factor scores, and (b) words as loadings on semantic factors. Every document is associated with a set of factor scores that act as coordinates within a semantic space created (i.e., "spanned") by the factors. The mathematical similarity between documents is calculated as the distance between them in the shared semantic space, through calculating the angle between the vectors that give the coordinates of two documents ("cosine similarity," Charikar, 2002).

This method has led to a number of successful uses of LSA in education contexts. For example, student responses on a test can be automatically scored by calculating the distance of their response from an ideal response in the semantic space (e.g., Wolfe & Goldman, 2003). Landauer and Dumais (1997) built an LSA model on a schoolbook corpus, and used LSA to measure the distance between the text of the test questions and the text of the multiple choice answer choices; they found the closest answer to be correct in 64.4% of the cases. Campbell and Pennebaker (2003) used LSA to measure changes in the use of language across writing sessions about traumatic events, to see if changes in writing style or content were associated with fewer hospital visits. They used LSA to create different semantic spaces for function (prepositions, pronouns, etc.) and content words, and showed that only changes in the function (predominantly pronouns) space predicted better health outcomes. In other words, among those who were asked to write about emotional trauma, the less similar (and more different) the essays were in their use of pronouns, the bigger the positive health effects.

Though LSA offers a robust method to quantify semantic differences between documents, the interpretability of its semantic factors is limited. Words negatively loading onto a factor are hard to interpret, and generally words loading onto the same LSA factor are not semantically coherent. In part, this shortcoming is a result of approximating language as a space: words have a number of relationships that are less symmetric than this assumption imposes. For example, *buckle* is semantically close to *belt*, *asteroid* is semantically close to *belt*, but *buckle* is not close to *asteroid* ("the triangle inequality," for a fuller discussion see Griffiths et al., 2007). Words vary tremendously in frequency (see Figure 1), which may significantly influence the prediction of associations between words: Given that *buckle* occurs more frequently than *asteroid*, the association between *buckle* and *belt* will greatly diminish the association between *asteroid* and *belt*. In short, LSA imposes constraints that the semantic structure of language cannot follow.

Latent Dirichlet Allocation (LDA). LDA, developed by Blei, Ng and Jordan (2003) is better suited than LSA to identifying commonalities across words and documents. It is less straightforward than LSA's factor analysis of the word-by-document matrix, but yields more interpretable factors. Like LSA, it uses the WBD matrix, encoding how words are distributed over documents. LDA assumes that the occurrence of words can be explained by unobserved groups, called topics.

Topics created ("modelled") through LDA are interpretable, semantically-coherent sets of words that occur in the same contexts. They can be thought of as data-driven "micro-dictionaries" in which words have weight, based on their contribution to the topic. This results in an elegant feature-reduction of the language space. For example, rather than the users' language being described as distributions over 20,000 words and phrases, they can be expressed as a distribution over a number of k topics, where k can be chosen freely. The resulting topics are often helpful in summarizing the content and semantic contexts of a given text corpus.

LDA assumes that each word can be attributed to one of the document's topics. The LDA algorithm considers which word belongs to which topic and which topics constitute a given document, and iterates until an optimal equilibrium is reached. This results in a set of posterior probability distributions, which approximates documents as distributions over topics, and topics as probability distributions over words (see Figure 2).

Unlike LSA, the topics are semantically coherent. Words that co-occur in the same contexts are combined, and words only load positively onto topics. Through this "structured representation," LDA can take different word senses into account: *belt* will appear with *asteroid* in an astronomy topic, as those words were observed to co-occur in some documents. A separate topic will include *belt* and *buckle* and other clothing items. Thus, different senses of a word are cleanly separated. A word is seen within the context of the other topic words with which it co-occurs. Further, differences in word frequency is no longer problematic, as the word senses are treated separately. As such, the topic modelling process generates topic units of analysis which overcome word sense ambiguities, one of the major sources of potential confusion with the top-down dictionary-based approach.

Topic modelling vs. extraction. Importantly, the generation of topics ("topic modelling") and their application ("topic extraction") of the previously modelled topics are two different processes that need not be based on the same dataset ("corpus"). That is, one set of data can be used to develop the topics, and then the topics can be used as data-driven dictionaries in a second dataset. In fact, larger datasets results in more fine-grained, semantically coherent, and “cleaner” topics, thus it is often preferable to model one’s topics on a larger language sample than may be analyzed in a given study. As topic modelling works best on larger sets of documents, a large corpus can be used to model topics of high quality and semantic coherence, which can then be applied to smaller datasets, effectively leveraging the language information contained in the larger dataset for building the variables to be explored in a smaller dataset. Since its introduction in 2003, modifying and extending the original LDA model to better address different applications has become its own research area (e.g., Blei, 2012). Atkins et al. (2012) provide an excellent worked example of the application of LDA in the psychology literature.

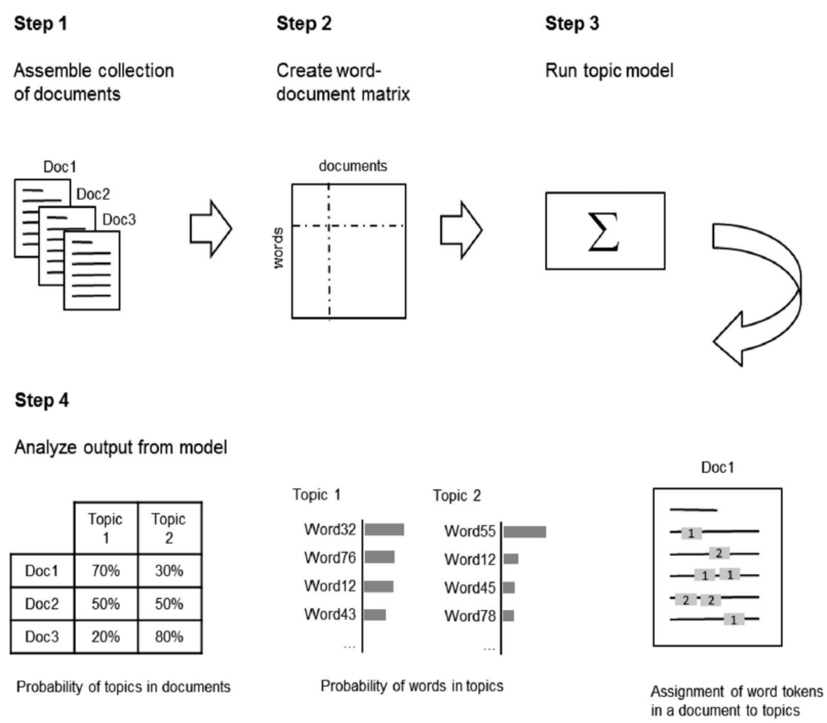


Figure 2. The process of topic modelling using LDA. Documents are collected (step 1) and represented as a word-document matric (WDM, step 2). Topic models are run on the WDM (step 3). The two sets of probability distributions (probability of topics in documents and probability of words in topics) are then fit simultaneously (Step 4), based

on assigning individual word occurrences in documents to topics. Adapted from Griffiths et al., 2007 with permission.

Differential Language Analysis. We have proposed Differential Language Analysis (DLA) as a method for conducting exploratory open-vocabulary analyses for a given variable (Schwartz et al., 2013b; Kern et al., 2014). In this fairly straight-forward approach, every word (or 1-gram) is individually correlated against an outcome. For example, if language samples are available for 1,000 people for whom self-reported extraversion scores are also known, for a given word we derive the its 1,000 relative frequencies and correlate them with the 1,000 extraversion scores. This provides a single correlation coefficient for a word (for example, the word "party" might be correlated with extraversion at r = .23 across 1,000 individuals).

This procedure is repeated for all words in the vocabulary, and other "tokens"--other separable pieces of text like emoticons ("*:-)*", "^.^") or punctuations (*!!!!*)--as well as phrases of up to 3 tokens ("1-to-3-grams"). Once the relative frequency of all 1-to-3-grams has been individually correlated against an outcome, the most positively and negatively correlated words and phrases can be shortlisted for an outcome, yielding the words that most *differentiate* an outcome. If a dataset is sufficiently large, even very rare words in the long tail of the Zipfian distribution can be suitable units of analysis. (For a full overview of the method, see Schwartz et al., 2013b. For examples of DLA applied to personality, age, and gender, see Kern et al., 2014a; Kern et al., 2014b, and Park et al., 2016 respectively.)

The Need for a Quantitative Comparison

Currently the most common approach to text analysis in psychology is through closed-vocabulary methods. With over 2,100 citations, LIWC is by far the most popular computerized text analysis program used in psychology. However, GI, DICTION, and LIWC have never been directly compared in their ability to generate psychological insight from text. By testing the three programs across the same dataset, their respective strengths and weaknesses can be illuminated.

Further, with the increasing availability of computational power, methods like topic modelling promise to capture markedly more conceptual and behavioral nuances than the closed-vocabulary methods. While LIWC has been cited several thousand times, as of March 2017, the key LSA publications (Deerwester et al., 1990; Landauer & Dumais, 1997; Landauer, Foltz, & Laham 1998) have received 18,500 citations in the computational disciplines, and the publication that introduced LDA (Blei, Ng, & Jordan, 2003) has been cited 13,000 times.

With the recent availability of vast amounts of digital text, or “big data” (Gandomi & Haider, 2015; Kosinski, Matz, Gosling, Popov, & Stillwell, 2015; Manyika et al., 2011), data that capture users’ behavior on the web are increasingly available, through sources such as online forums (e.g., Gross & Murthy, 2014), search queries (e.g., Brownstein, Freifeld, & Madoff, 2009), and social media datasets (e.g., Fan, Zhao, Chen, & Xu, 2014; McKelvey, DiGrazia, & Rojas, 2014; Spertus, Sahami, & Buyukkokten, 2005; Youyou, Kosinski, & Stillwell, 2015; Yu & Wang, 2015). Such datasets potentially will play a role in the future of psychological science, but their utility depends on the ability to make sense of the data. Figure 3 documents the growing number of publications on Facebook and Twitter. The question of how to best analyze this new generation of datasets is important and timely. Guidance is needed as to which text analysis program is most appropriate for a text dataset of a given size, and what value might be added by using open-vocabulary methods.

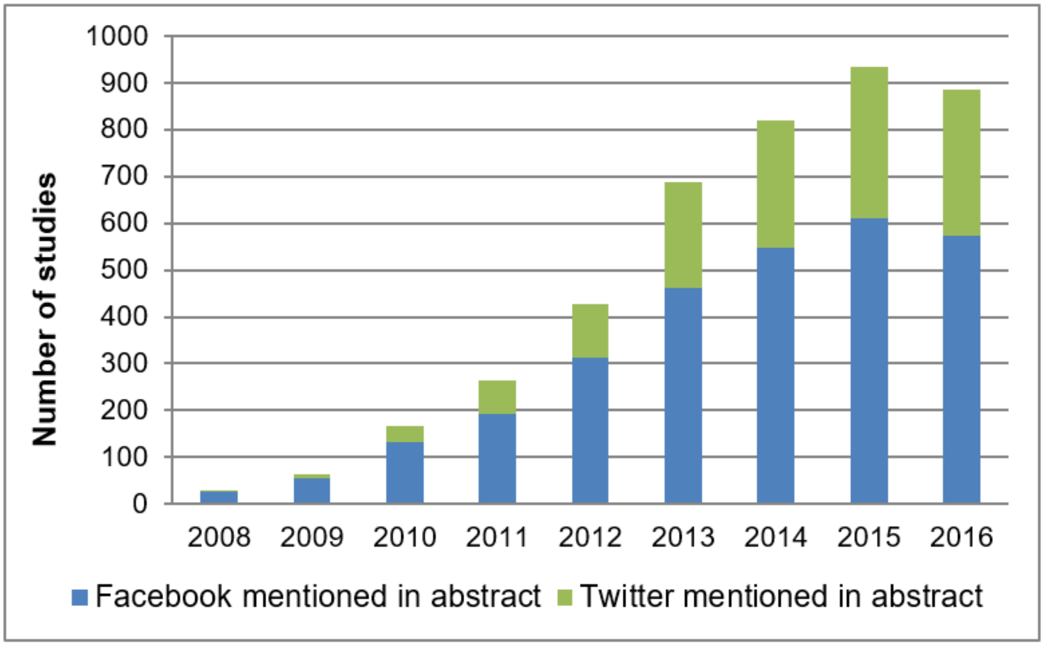


Figure 3. Number of studies indexed by PsycINFO mentioning Facebook (blue) or Twitter (green) in the abstract from 2008 to 2016 (as of March 2017, 2016 indexing not complete).

The Present Study

This study aims to provide a comprehensive quantitative comparison amongst the leading closed and open-vocabulary methods for language analysis, to empirically inform best practice approaches. We use one of the most popular big social media datasets used by psychologists, the MyPersonality dataset (Kosinski et al., 2013), which includes text data from Facebook (www.facebook.com) as well as self-reported information. We apply the three most frequently used closed-vocabulary analysis programs and two open-vocabulary approaches that have recently been introduced to the psychological literature.

We discuss areas of overlap among the programs, and compare their ability to detect and validly capture psychological correlates of gender, age and Big-5 personality. In secondary analyses, we determine the sample sizes of social media users needed for exploratory language analyses using closed and open-vocabulary methods, and determine what number of LDA topics to extract.

Method

Survey and Demographic Data

The myPersonality Facebook dataset used in this study is the most popular social media dataset that has been used in psychology (e.g., Kosinski et al., 2013; Park et al., 2014; Schwartz et al., 2013; Wilmot et al., 2015; Youyou et al., 2015). MyPersonality was a third-party application on the Facebook platform installed by roughly 4.5 million users between 2007 and 2012 (Kosinski & Stillwell, 2012; Stillwell & Kosinski, 2004). The application allowed users to take psychological inventories and share their results with friends. Users completed 20 items from the International Personality Item Pool (IPIP; Goldberg et al., 2006), which assessed personality based on Costa and McCrae's (1992) five-factor model (Big Five). Personality is classified based on five factors: agreeableness (e.g., trusting, generous), conscientiousness (e.g., self-controlled, responsible), extraversion (e.g., outgoing, talkative), neuroticism (e.g., anxious, depressed), and openness (e.g., intellectual, artistic, insightful). All users agreed to the anonymous use of their survey responses for research purposes. Users also reported their age and gender (forced binary choice) as part of their Facebook profile; we limited the dataset to those users between 16 and 60 years. Mean user age was 24.57 years (median 21.00, *SD* 9.01), and over half (62.07%) were female.

Language Data

A subset of the users allowed the myPersonality application to access their Facebook status messages, which are undirected updates about the self which users post on their profile. These do not include messages between users, or comments on other users' statuses. We limited the sample to 65,896 individuals who in addition to having reported age, gender and taken the personality survey also had at least 1,000 words across their status updates between January 2009 and November 2011, totaling over 12.722 million messages (see Kern et al., 2016, for discussion on word limits). Users wrote an average of 4,104 words across all status messages (median = 2,875, SD = 3,894, range = 1,000 to 82,538).

Linguistic Feature Extraction

We transformed each user's collection of status messages into numerical variables that capture the relative frequencies of three different sets of language features: (a) words and phrases, (b) dictionaries, and (c) LDA topics.

Words. The first step in text processing is to split users' statuses into tokens (i.e., single "words"). Tokens include single words, but also punctuation, non-conventional usages and spellings (e.g., *omg*, *wtf*) and emoticons (e.g., *:-]*, ^.^), which are common on social media. We used a social-media-appropriate tokenizer (happierfuntokenizing; Potts, 2011). We divided the frequencies of use for all tokens by a user's total number of tokens, yielding the users' relative frequencies of use.

Social media vocabularies tend be about one order of magnitude larger than the language used in transcripts (e.g., Atkins et al., 2012), as it includes many idiosyncratic misspellings, plays on words, and borrowings from other languages (e.g., zumbaaaaaaaaaaaaaaa, zombieapocalypse,). Thus, it is common to restrict analyses to words used by at least a certain fraction of the sample (e.g., Atkins et al., 2012). Accordingly, when using words as units of analyses in Differential Language Analyses, we limit the analysis to tokens that were used by at least 5% of the users (reducing the total number of distinct tokens (1-grams) from 1,680,708 to 2,986).

Dictionaries. Once word frequencies have been extracted for a given user, the words can be matched against existing dictionaries. Using our own Python codebase and MySQL infrastructure (see http://dlatk.wwbp.org), we extracted relative dictionary scores for the 73 dictionaries provided by LIWC, and 182 dictionaries provided by the General Inquirer. Wildcards were included, as dictated by the dictionaries (e.g., *happ** matches *happy* and *happier*). LIWC 2015 also generates "summary language variables" (a*nalytic thinking, clout, authentic, emotional tone*) which combine the relative frequencies of other dictionaries. So as not to miss these summary variables when considering LIWC's associations with demographics and personality, we used LIWC2015's batch mode to extract these in conjunction with the dictionary frequencies. These scores were then fed back into our database infrastructure for subsequent analysis.

Similarly, DICTION creates five *master* variables that combine 31 dictionary scores as well as nine language statistics. To obtain these master variables, we exported all the Facebook statuses, and ran them through DICTION's batch mode in combinations of about 3,000 users at a time, yielding a score for all 45 DICTION variables for each user, and imported back into our MySQL/Python analysis pipeline.

Although the GI's original 1960s implementations included rule-based routines to disambiguate words and account for their order, we limited calculations to the relative frequencies of dictionaries. We believe that future users are more likely to use the dictionaries in a general-purpose word-counting software implementation, such as LIWC or our python code base.

Phrases. The extraction of words (single tokens) and dictionaries disregards the order of words, treating all words as equal. Extracting phrases (in this case, sequences of two [2-grams] and three tokens [3-grams]) can capture distinctive language expressions that would otherwise be lost (e.g., *thank you*, *happy birthday*, *can't wait*). Rather than consider all possible combinations of two or three words that appear in a corpus, it is reasonable to consider only phrases which appear with higher probability (relative frequency) than the independent probabilities of their constituent words would suggest. For example, the phrase *happy birthday* appears with higher probability than the independent probabilities of *happy* and *birthday* would suggest; if *happy birthday* were not a special phrase, it would only be about as common as *great birthday*, rather than 10 times more likely. We used the pointwise mutual information (PMI) to quantify these probabilities, keeping phrases with a threshold above 3. A PMI threshold of three would mean that for inclusion in the analysis, a phrase would have to appear three times as often as the relative frequencies of its constituent words would suggest (for a full discussion, see Kern et al., in press and Schwartz et al. 2013b).

Phrase frequencies were divided by the user's total number of words, yielding relative frequencies. We again kept the 11,894 phrases (1-to-3-grams) that were used at least by 5% of the users.

Topic extraction. For our main analysis, we used a previously modelled set of 2000 Facebook topics, applying the existing topics to the current dataset. The topics were originally modeled using 14 million Facebook statuses (Schwartz et al., 2013b), and have been applied in subsequent studies to Facebook (e.g., Kern et al., 2014; Kern et al., 2014b; Park et al., 2014) and Twitter language data (Schwartz et al, 2013a; Eichstaedt et al., 2015) (The topics can be downloaded on http://wwbp.org/data, akin to weighted micro-dictionaries).

Given a set of documents (in our case, Facebook statuses), the LDA topic modelling process seeks to describe the documents as a combination of a small number of topics, which in turn are constituted by a small number of words. As shown in Figure 2, LDA creates a distributions of weights ("posterior probabilities") which capture how words are distributed in topics (*p(topic|word))* and how topics are distributed in documents (*p(topic|document*).Once topics are extracted, they can be used to describe the language used by a given unit of analysis (here, a Facebook user). We extracted the 2,000 previously modelled topics from the language of every Facebook user in our dataset. We multiplied the word-topic weights (*p(topic|word)* which were determined during the modelling process with the relative frequencies of a users' words (*p(word|user)*), yielding the user's overall use of a given topic, *p(topic|user)=* $\sum_{words \in topic} p(topic|word) * p(word|user)$. Each user thus received 2,000 topic scores. We show the topics most correlated with age, gender and the Big-Five personality traits alongside the dictionary associations.

Primary Data Analyses

Our primary analyses involve correlational analyses across dictionaries, words, phrases, and topics. Regression analyses compare the predictive validity of the three programs and LDA topics. In addition, in a supplementary analysis we consider power and the impact of extracting different numbers of topics.

Correlational analyses. We first regressed each dictionary within the three closed-vocabulary programs against gender, age and Big Five personality. Next, we regressed the 11,894 words and phrases and 2,000 topics independently against those outcomes (running 13,984 separate regressions). Gender was entered as a covariate when regressing language variables against controlled for in the age regressions; age was controlled for gender regressions, and both age and gender were controlled for personality correlations.

Controlling for multiple comparisons**.** Given the large number of regressions, we used the Benjamini-Hochberg procedure (BH; Benjamini & Hochberg, 1995) to adjust the significance threshold based on the number of hypotheses being tested. That is, when correlating a set of features (such as the 73 LIWC dictionaries or 2,000 topics) with a given outcome, we corrected the customary significance threshold for the number of features that were simultaneously being correlated. The BH procedure is less conservative but more powerful than corrections of the family-wise error rate (like the Bonferroni correction; Holm, 1979), providing a balance between over and under-estimating potential effects.

Word clouds (words and phrases)**.** We have found word clouds to be a space-efficient way to visualize the most highly correlated 50 words and phrases. Traditional word clouds used to summarize text (e.g., www.wordle.net) scale words by frequency of occurrence. Although this encodes direct frequencies, this approach does not visualize differences between groups or traits. Instead, we use our Python codebase (see wwbp.org/data) to generate word clouds that scale the words by the magnitude of their correlation coefficient, with larger words indicating a stronger (positive or negative) correlation with the outcome. Word color is used to capture frequency, from red (frequently used) to blue (moderately used) to grey (rarely used). In this way, the word clouds summarize the words and phrases that most discriminate a given outcome while still allowing the reader to keep track of frequency. In addition, we prune duplicate mentions of a word (i.e., when a single word also occurs in a phrase), giving preference to more highly correlated phrases over single words (explained in more depth in Schwartz et al., 2013.)

Topic word clouds. We visualize topics as word clouds that show the 10 words with the largest prevalence in the topic (that is, product of overall word frequency and word weight in a given topic [$p(topic, word) = p(topic|word) * p(word)$]), with the size of the words scaled by descending prevalence, such that the largest word has the highest prevalence in the topic. We show the eight topics with the strongest associations. On occasion, the LDA algorithm creates topics that are very similar to one another (duplicates); we excluded a topic for visualization if it shared more than 25% of its top 15 words with the top 15 words of a more strongly correlated topic.

Prediction. To quantify the amount of outcome-related variance captured by the dictionaries and topics, we separately used each set of dictionaries and the 2,000 topics as features predicting each outcome (gender, age, and Big Five personality traits). In choosing the prediction models, our goal was not necessarily to reach state of the art prediction performances (cf. Park et al., 2014; Sap et al., 2014; Schwartz et al., 2013b), but use a type of prediction model that would be appropriate for both a relative small (e.g., 36 DICTION dictionaries) and large number of features (e.g., 2,000 LDA topics).

We used penalized logistic regression (Gilbert, 2012) for the binary gender variable and penalized regression (or ridge regression; Hoerl & Kennard, 1970) for the continuous age and personality variables. Both techniques are fairly straight-forward machine learning extensions of logistic regression and linear regression, in which the squared magnitude of the coefficients is added as a penalty to the error term, and this penalized error and the squared error are minimized simultaneously when fitting the coefficients. This biases the coefficients towards zero, addressing problems of colinearity between the coefficients (language features are often highly intercorrelated) and reducing overfitting, thereby increasing the ability of the fitted model to generalize to new data (Fan, Chang, Hsieh, Wang, & Lin, 2008). The relative importance of the squared error and the penalization term during the model fitting is controlled by a "hyperparameter" that is chosen automatically during the model fitting.

We report ten-fold cross-validated prediction accuracies. The data are split randomly into ten random subsets ("folds"), and a model is fit over nine of the folds ("training set"). The trained model is then applied to the remaining fold ("test set"), and its predicted outcome values (e.g., user extraversion scores) are compared to the actual values in the test set. Accuracy is calculated as the Pearson correlation between the predicted and actual outcome values. This procedure is then repeated in round-robin fashion until every fold has been the test set once. The final predictive accuracy is the average of the ten accuracies.

Secondary Data Analysis

When carrying out open-vocabulary language analyses, the researcher needs to make a number of decisions, including if the data set is of sufficient size and has a sufficient amount of language per unit of observation (e.g., word count per user) to yield sufficient power for an exploratory analysis given different sets of language-derived variables, and if topics are extracted, how many topics should be extracted.

Power analyses: number of users. A possible advantage of dictionary-based methods is their relatively smaller number of language features (Diction: 36, LIWC: 73, General Inquirer: 182), increasing their power when using associations with language features as an exploratory method (while controlling for multiple comparisons). To inform which method is appropriate for datasets of different sizes, we correlated the different sets of language features with age and gender and the personality dimensions across randomly-selected samples of 50, 500, 1,000, 2,000, 5,000, 15,000 and 50,000 users. We used the Benjamini-Hochberg method to correct for multiple comparisons.

Choosing the number of topics to extract. The key parameter that needs to be chosen during the topic modeling process is the numbers of topics to extract (k). We previously found that given a large enough dataset, extracting more topics creates topics that have more specificity, at the cost of some topics being very similar (Kern et al., 2016). To explore the choice of different numbers of topics, we used LDA to model different number of topics (50, 500 and 2,000 topics) across the Facebook dataset with different numbers of Facebook statuses (50, 500, 5000, 50,000, 500,000 and 5 million statuses), yielding a total of 18 different sets of topics (3 choices for number of topics x 6 different datasets with different number of statuses). We examined the ability of the 50, 500, 2000 topics modeled over 5 million statuses to distinguish contexts and word-senses of the word *play*. To quantify the information captured by the different number of topics, we first used the 18 different sets of extracted topic frequencies as features in 18 machine learning prediction models (ridge-regression), predicting the age, gender, and Big Five personality of the users, and report the average out-of sample (cross-validated) prediction accuracies.

Results

GI, DICTION, and LIWC overlap in their coverage of some concepts, while each program includes unique dictionaries. All three programs include dictionaries for positive affect, negative affect, and first person singular pronouns. Other concepts that are covered in dictionaries across the programs include cognition and complexity of language (Harvard-IV *abstract vocabulary*; DICTION *cognition*; LIWC *insight, tentative, causation, cognitive processes*; Lasswell *enlightenment* dictionaries,), economic and fiscal concerns (Harvard-IV *economic*; Lasswell *wealth* dictionaries; LIWC *money, work, achievement;*,).

Table 1 shows the intercorrelations across 65,896 Facebook users. For the affect dimensions, GI and LIWC show larger intercorrelations with one another than with DICTION. Due to LIWC's hierarchical structure, sub-categories often correlate highly with their respective categories (e.g., the *first person singular* dictionary correlates at r = .77 with the overall *pronoun* dictionary).

Correlations between the dictionaries are mostly driven by overlap in the words that they contain. A few very frequent words often contribute the majority of counts in dictionaries; when they occur in multiple dictionaries, these dictionaries will be highly correlated.

Table 1

Intercorrelations Amongst Positive Affect, Negative Affect, and Pronoun Dictionaries.

	General Inquirer			Diction		LIWC
	Lasswell	Harvard IV	Osgood			
	Positive Affect	Pleasure	Positive	Optimism	Satisfaction	Affect
General Inquirer						
Pleasure	.48					
Positive	.70	.63				
Diction						
Optimism	.33	.45	.33			
Satisfaction	.31	.53	.34	.72		
LIWC						
Affect	.37	.47	.33	.27	.37	
Pos. Emotion	.45	.60	.42	.46	.45	.85

	General Inquirer				Diction		LIWC	
	Lasswell	Harvard IV	Stanford					
	Negative Affect	Vice	Negative	Hostile	Hardship	Blame	Swear	Negative Emotion
General Inquirer								
Vice	.59							
Negative	.68	.76						
Hostile	.60	.54	.85					
Diction								
Hardship	.26	.23	.26	.17				
Blame	.27	.27	.22	.14	.12			
LIWC								
Swear	.39	.26	.38	.37	.13	.10		
Negative Emotion	.56	.45	.49	.34	.36	.28	.61	
Anger	.48	.37	.46	.41	.24	.17	.87	.76

	Gen. Inquirer	Diction	LIWC	
	Harvard IV: Self	Self-reference	Pronouns	Pers. pronouns
Diction				
Self-reference	.75			
LIWC				
Pronouns	.75	.49		
Pers. Pronouns	.70	.60	.96	
1st. pers. sing.	.92	.80	.75	.77

Table 2

Standardized Regression Coefficients between User Gender and Different Dictionaries across 65,896 Facebook users (controlled for Age)

	General Inquirer						DICTION		Linguistic Inquiry and Word Count (LIWC 2015)			
	Lasswell		Harvard IV		Stanford				LIWC (other)		LIWC (psych. processes)	
	Dictionary	β	Dictionary	β	Dictionary	β	Dictionary	β	Dictionary	β	Dictionary	β
Female	Affect		Pleasure	.29	Affiliation	.12	Optimism (m)	.14	Emotional tone (m)	.27	Social processes	.12
	Affect-Other	.28	Females	.28	Passive	.09	+Satisfaction	.22	Personal pronoun	.17	Female reference	.30
	Affect-Domain	.21	Emotion	.25	Positive	.09	+Praise	.08	1st pers singular	.16	Family	.28
	Affect-Gain	.16	Kinship	.20	Weak	.06	+Inspiration	.05	3rd pers singular	.11	Affective process	.25
	Affect-Participants	.05	Self	.15	Submit	.05	-Blame	.04	2nd person	.07	Positive emotion	.29
	Wellbeing-Total	.15	Children	.15			Certainty (m)		Total pronouns	.11	Home	.21
	Wellbeing-Psych.	.24	Independent Adj.	.12			+Insistence	.07	Common adverbs	.09	Netspeak	.18
	Wellbeing-Participants	.16	State Verb	.12			-Self-reference	.15	Common verbs	.07	Affiliation	.17
	Positive-Affect	.11	Need	.11			+Tenacity	.06	Conjunctions	.07	Future focus	.10
	Transaction-Gain	.10	Evaluation 2	.10			Human Interest	.12	Common adjectives	.06	Nonfluencies	.10
	Respect-Lose	.07					Temporal	.05				
Male	Wealth-Total	.19	Military	.21	Strength	.09	Realism		Articles	.24	Death	.22
	Wealth-Other	.19	Movement-Exert	.21	Hostile	.08	+Familiarity	.09	Analytical thinking (m)	.19	Anger	.21
	Power-Total	.18	Political	.19	Negative	.07	+Spatial	.09	Comparisons	.12	Drives	
	Power-Arenas	.15	Economic	.16	Understated	.06	-Complexity	.08	Prepositions	.12	Power	.20
	Power-Conflict	.14	Region	.15	Active	.06	Activity		Impersonal pronouns	.08	Achievement	.13
	Power-Participants	.14	Space	.15	Power	.06	+Aggression	.10	Quantifiers	.06	Risk	.09
	(Ordinary)		Doctrine	.15			+Accomplishment	.07	Interrogatives	.06	Swear words	.19
	Power-Authority	.13	Abstract vocab.	.14			+Communication	.07	3rd pers plural	.05	Sexual	.19
	Power-Loss	.12	Collectives	.14			Commonality		Numbers	.04	Space	.16
	Arenas	.17	Expressive	.13			+Centrality	.08			Money	.11
	Religion	.14					-Diversity	.06			Tentative	.09
							-Exclusion	.05				
							Collectives	.06				

Note. All coefficients are significant at $p < .001$, corrected for multiple comparisons. (m) designates “master” categories that combine frequencies of multiple dictionaries.

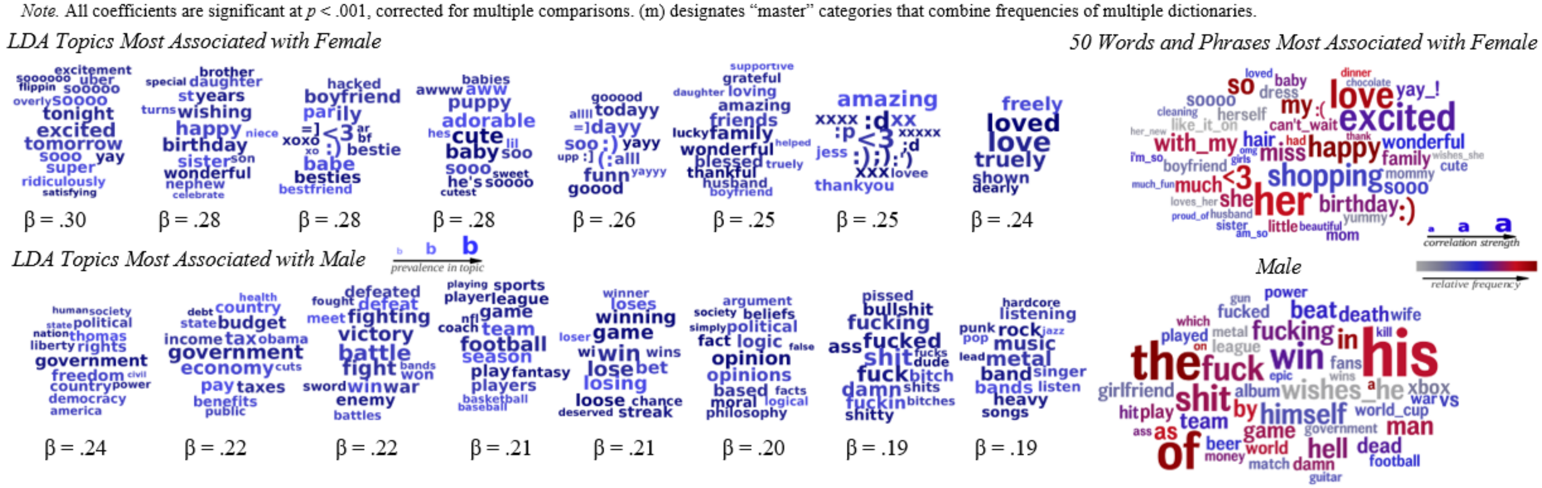


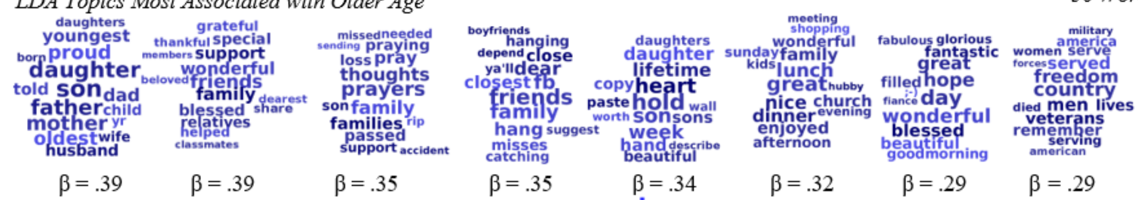
Table 3

Standardized Regression Coefficients between User Age and different Dictionaries across 65,896 Facebook users (controlled for Gender)

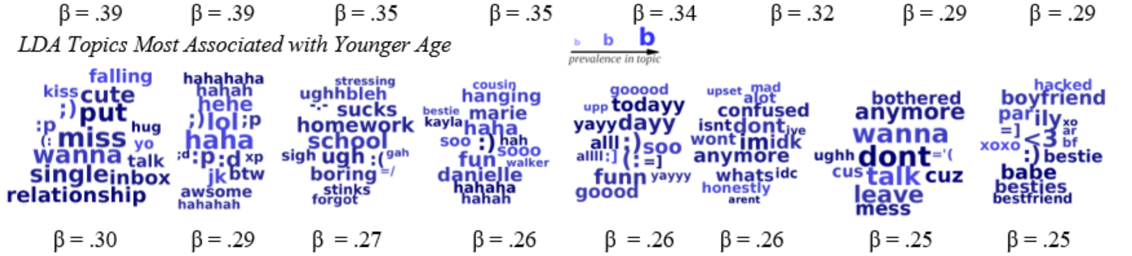
	General Inquirer						DICTION		Linguistic Inquiry and Word Count (LIWC 2015)			
	Lasswell		Harvard IV		Stanford				LIWC (other)		LIWC (psych. processes)	
	Dictionary	β	Dictionary	β	Dictionary	β	Dictionary	β	Dictionary	β	Dictionary	β
Older	Power-Total	.17	Kinship	.29	Power	.16	Realism	.16	Clout (m)	.29	Social processes	.19
	Power-Other	.23	Economic	.25	Positive	.12	+Familiarity	.24	Articles	.29	Family	.27
	Power-Participants	.17	Communication Tools	.24	Affiliation	.11	+Human Interest	.21	Prepositions	.28	Drives	
	(Authority)		Human	.21	Submit	.09	-Complexity	.11	Quantifiers	.24	Affiliation	.21
	Wealth-Total	.22	1st pers. plural	.20	Strength	.04	Certainty	.23	Emotional tone(m)	.21	Power	.20
	Wealth-Other	.19	Political	.18	Understated	.04	+Collectives	.11	Analytical thinking (m)	.21	Relativity	.14
	Transaction-Gain	.19	Region	.18	Overstated	.02	+Insistence	.10	Personal pronouns		Space	.21
	Respect-Other	.20	Role	.17			+Tenacity	.09	3rd pers plural	.24	Personal concerns	
	Means	.18	Objects	.17			Rapport	.16	1st pers plural	.18	Money	.20
	Affect-Participants	.18	Male	.16			Optimism	.12	3rd pers singular	.13	Religion	.18
	Wellbeing-Gain	.16							Function words	.13	Home	.17
Younger	Negative-Affect	.24	Self	.20	Negative	.19	Certainty		Personal pronouns	.14	Affective process	.20
	Affect-Gain	.18	Academic vocab.	.19	Hostile	.16	-Self-reference	.22	1st pers singular.	.27	Negative emotion	.33
	Wellbeing-Loss	.17	Emotion	.16	Passive		-Ambivalence	.03	Negations	.18	Anger	.27
	Rectitude-Gains	.12	Pain	.14			-Variety	.05	Common Adverbs	.17	Sadness	.17
	Enlightenm.-Ends	.12	Disagreement	.14			Optimism		Pronouns	.08	Informal language	
	Transaction-Loss	.09	Vice	.14			-Hardship	.12	Authentic(m)	.07	Netspeak	.30
	Power-Conflict	.08	Expressive	.12			-Blame	.11	Numbers	.05	Swear words	.21
	Affect-Loss	.07	Nature Process	.11			-Denial	.04			Assent	.15
	Enlightenm.-Other	.06	Say	.10			Present-Concern	.03			Nonfluencies	.14
	Denial	.06	Very	.09			Activity				Biological process	
							-Cognition	.03			Body	.17
							+Aggression	.03			Sexual	.16
							+Motion	.02				

Note. All coefficients are significant at p < .001, corrected for multiple comparisons. (m) designates "master" categories that combine frequencies of multiple dictionaries.

LDA Topics Most Associated with Older Age



LDA Topics Most Associated with Younger Age



50 Words and Phrases Most Associated with Older Age

Younger Age

Regression Analyses

We first examined associations between the three dictionaries and gender, age, and Big Five personality. We report the highest standardized regression coefficients between the dictionaries and outcomes;[2] as well as the most associated topics (from a set of 2,000 topics) and words and phrases.

Gender. As seen in Table 2, across programs, being female was associated with with dictionaries capturing positive emotion (GI-Lasswell: *affect-other*, β = .28, *well-being psychological*, β = .24; GI Harvard-IV: *pleasure*, β = .29, *emotion*, β = .25; GI-Stanford: *positive*, β = .09; LIWC: *positive emotion*, β = .29) and first person pronouns (GI-Harvard-IV: *self*, β = .15, DICTION: *self-reference*, β = .15; LIWC: *first person singular*, β = .16). This consistency across sets of dictionaries is not surprising given the moderate-to-high intercorrelations between these dictionaries (c.f. Table 1). The GI *female* and LIWC *female references* dictionaries showed some of the strongest associations with female gender (β = .28 and β = .30, respectively). These dictionaries contains both female nouns (*girl, mom*) as well as female pronouns (*her, she)*. Similarly, female users used more language associated with close relationships (GI-Harvard-IV: *kinship*, β = .20; GI-Stanford: *affiliation*, β = .12; LIWC: *family*, β = .28, *friends*, β = .09), aligning with prior findings that women use more socially oriented words than men (Pennebaker, 2011).

[2] When reporting dictionary correlations across Tables 2-8, we take into account that dictionaries exhibit a hierarchical structure (e.g., words in the LIWC *anger* dictionary are part of the LIWC *negative emotion* dictionary). In cases in which the broader dictionary category showed a significant association, sub-dictionaries within that category are placed below it. For cases in which the superordinate dictionary category did not show a significant association, the higher order dictionary was included without a regression coefficient if two or more of its sub-dictionaries were significantly associated.

By contrast, being male was associated with dictionaries reflecting negative emotion (GI-Stanford: *negative*, β = .07; LIWC: *negative emotion*, β = .02, *swear*, β = .19), economic concerns (GI-Lasswell: *wealth-total*, β = .19; GI-Harvard-IV: *economic*, β = .16; LIWC: *money*, β = .11), and hostility and aggression (GI-Harvard-IV: *military*, β = .21, *political*, β = .19; GI-Stanford: *hostile*, β = .08, *strength*, β = .09; DICTION: *aggression*, β = .10). The GI-Stanford dictionaries clearly separate the genders along the *affiliative-passive-positive* (female) and *hostile-strength-negative* (male) dimensions.

While the closed-vocabulary approaches suggest that language indicating positive emotion language tends to be associated with women, the DLA word clouds reveal which emotions in particular show the strongest associations; they tend to be high-arousal emotions (*excited*, *happy*, yay!) and mentions of *love*.

The LDA topics reveal that words indicating economic concerns often appear in the context of political-fiscal debate, such as *tax*, *budget*, *economy, government*, *income,* and *benefits* (topic association β = .22). The LDA topics suggest that language associations around hostility and aggression may in large part be specifically driven by competition (*battle, victory, fight,* topic association β = .22), political debate (*country, power*, β = .24), as well as sports (*win, lose, bet*, β = .21).

Being male was also associated with the use of articles and prepositions suggestive of higher object orientation and noun use, born out in both the LIWC *articles* (*r* = .24) and *prepositions* (β = .12) dictionaries, as well as the most-associated open-vocabulary words (*of*, *the*, *in*, *by*).

Age. As Table 3 shows, younger age was associated with self-reference (GI-Harvard-IV: *self*, β = .20; DICTION: *self-reference*, β = .22; LIWC: *first person singular,*

$\beta = .27$) and negative emotion (GI-Lasswell: *negative affect*, $\beta = .24$; GI-Stanford: *negative*, $\beta = .19$; LIWC: *negative emotion*, $\beta = .33$; *swear*, $\beta = .21$). Conversely, older age was associated with talking about others (LIWC: *third person plural (they)*, $\beta = .24$, *first person plural (we)*, $\beta = .18$, *third person singular (s/he)*, $\beta = .13$), economic concerns (GI-Lasswell: *wealth-total*, $\beta = .22$; GI-Harvard-IV: *economic*, $\beta = .25$; LIWC: *money*, $\beta = .20$), and family and social categories (GI-Lasswell: *Respect-Other*, $\beta = .20$; GI-Harvard-IV: *kinship*, $\beta = .29$, LIWC: *family*, $\beta = .27$).

LDA topics mirrored these themes, with friends and family topics (*daughter, son, father, mother)* being the most strongly associated with older age ($\beta = .39$). The DLA word clouds mark younger age by the use of emoticons and symbols (*<3, :(, :), :d*), colloquialisms and contractions (*wanna*, *kinda, cant, im*), and suggest *hate, bored,* and *stupid* as specific expressions of negative emotions. Language of older individuals showed markers of longer sentences and increased use of nouns (LIWC: *articles*, $r = .29$, *prepositions*, $r = .28$), which was mirrored in the DLA findings (*the, of, for*).

Table 4

Standardized Regression Coefficients between Agreeableness and different Dictionaries across 65,896 Facebook users (controlled for Age and Gender)

	General Inquirer						DICTION		Linguistic Inquiry and Word Count (LIWC 2015)			
	Lasswell		Harvard IV		Stanford				LIWC (other)		LIWC (psych. processes)	
	Dictionary	β	Dictionary	β	Dictionary	β	Dictionary	β	Dictionary	β	Dictionary	β
More Agr.	Wellbeing-Psychological	.08	Pleasure	.12	Positive	.10	Optimism	.11	Emotional tone (m)	.21	Positive emotion	.14
	Affect-Other	.07	Time Broad	.08	Affiliation	.06	+Satisfaction	.08	Clout (m)	.07	Drives	.09
	Positive-Affect	.07	Religion	.07	Overstated	.05	+Praise	.07	Personal pronouns		Affiliation	.09
	Certainty	.06	1st pers. plural	.06	Power	.04	+Inspiration	.05	1st pers plural	.06	Reward	.06
	Space-Time	.06	Virtue	.06	Strength	.04	Certainty	.06	Common adjectives	.05	Achievement	.05
	Rectitude-Ends	.06	Expressive	.05	Submit	.03	+Leveling	.03	Prepositions	.04	Relativity	.07
	Transaction-Gain	.05	Quantity Ordinal	.04	Passive	.02	+Insistence	.06	Quantifiers	.04	Time	.08
	Respect-Total	.05	Names	.04	Understated	.02	Realism	.04	Authentic (m)	.03	Motion	.05
	Respect-Lose	.04	Independent Adj.	.04			+Temporal	.05			Future focus	.07
	Skill-Aesthetic	.04	Sky	.04							Religion	.07
Less Agr.	Negative-Affect	.09	Vice	.08	Negative	.08	Optimism		Negations	.06	Negative emotion	.15
	Wellbeing-Loss	.05	Disagreement	.06	Hostile	.06	-Hardship	.05	Personal pronouns		Anger	.20
	Denial	.04	Negation	.04			-Blame	.04	1st pers singular	.03	Anxiety	.03
	Power		Races	.03			-Denial	.03	3rd pers plural	.03	Swear words	.16
	Power-Authority	.03	Increase	.03			Aggression	.04			Biological process	.05
	Power-Participants	.03	Say	.03			Variety	.03			Sexual	.13
	(Authority)		Color	.03			Self-reference	.02			Body	.08
	Power-Participants	.03	Nature Process	.03			Communication	.02			Personal concerns	
	(Ordinary)		Body Parts	.03							Death	.10
	Negative-Value	.03	Movement-Exert	.03							Money	.04
	Affect-Loss	.03									Risk	.05
	Wealth-Transaction	.03										
	Skill-Participant	.02										

Note. All coefficients are significant at p < .001, corrected for multiple comparisons. (m) designates "master" categories that combine frequencies of multiple dictionaries.

LDA Topics Most Associated with Agreeableness

β = .12 β = .12 β = .11 β = .10 β = .10 β = .09 β = .09 β = .09

LDA Topics Most Associated with Agreeableness

prevalence in topic

β = .14 β = .09 β = .09 β = .09 β = .08 β = .07 β = .70 β = .07

50 Words and Phrases Most Correlated with

High Agreeableness

correlation strength

Lower Agreeableness

relative frequency

Table 5

Standardized Regression Coefficients between Conscientiousness and different Dictionaries across 65,896 Facebook users (controlled for Age and Gender)

	General Inquirer						DICTION		Linguistic Inquiry and Word Count (LIWC 2015)			
	Lasswell		Harvard IV		Stanford				LIWC (other)		LIWC (psych. processes)	
	Dictionary	β	Dictionary	β	Dictionary	β	Dictionary	β	Dictionary	β	Dictionary	β
More Con.	Affect-Other	.08	Time Broad	.10	Positive	.09	Certainty	.07	Emotional tone (m)	.17	Drives	.12
	Space-Time	.07	Pleasure	.09	Strength	.07	+Insistence	.09	Prepositions	.07	Achievement	.12
	Transaction-Gain	.07	Economic	.07	Affiliation	.07	+Collectives	.05	Clout (m)	.06	Reward	.09
	Certainty	.07	Strength	.07	Power	.05	Optimism	.08	Quantifiers	.06	Affiliation	.07
	Wellbeing-Psych.	.06	Travel	.07	Submit	.05	+Satisfaction	.05	Personal pronouns		Relativity	.10
	Skill-Other	.06	Virtue	.07	Overstated	.04	+Praise	.05	1st pers plural	.05	Time	.11
	Ends	.06	Comparison	.06	Active	.04	Realism	.07	Analytical thinking(m)	.05	Motion	.06
	Nations	.05	1st Person Plural	.06	Understated	.02	+Temporal	.07	Common adjectives	.05	Positive emotion	.11
	Skill-Total	.05	Interpretative	.06			+Familiarity	.05	Authentic (m)	.04	Work	.11
	Positive-Affect	.05	Action	.06			Accomplishment	.06	Articles	.03	Future focus	.07
Less Con.	Negative-Affect	.06	Vice	.07	Negative	.05	Certainty		Personal pronouns	.04	Negative emotion	.12
	Wellbeing-Loss	.05	Color	.05			-Self-reference	.05	1st pers singular	.06	Anger	.13
	Enlightenm-Ends	.05	Self	.04			-Variety	.04	Negations	.05	Sadness	.05
	Affect-Loss	.03	Disagreement	.04			-Ambivalence	.02			Biological process	.06
	Negative-Value	.02	Think	.04			Optimism				Sexual	.11
	Form	.02	Tool	.03			-Hardship	.04			Body	.09
			Races	.03			-Blame	.03			Swear words	.10
			Evaluation2	.03			-Denial	.02			Death	.09
			Nature Process	.03			Communication	.03			Perceptual process	.08
							Exclusion	.02			Hear	.07
							Past-Concern	.01				

Note. All coefficients are significant at p < .001, corrected for multiple comparisons. (m) designates "master" categories that combine frequencies scores of multiple dictionaries.

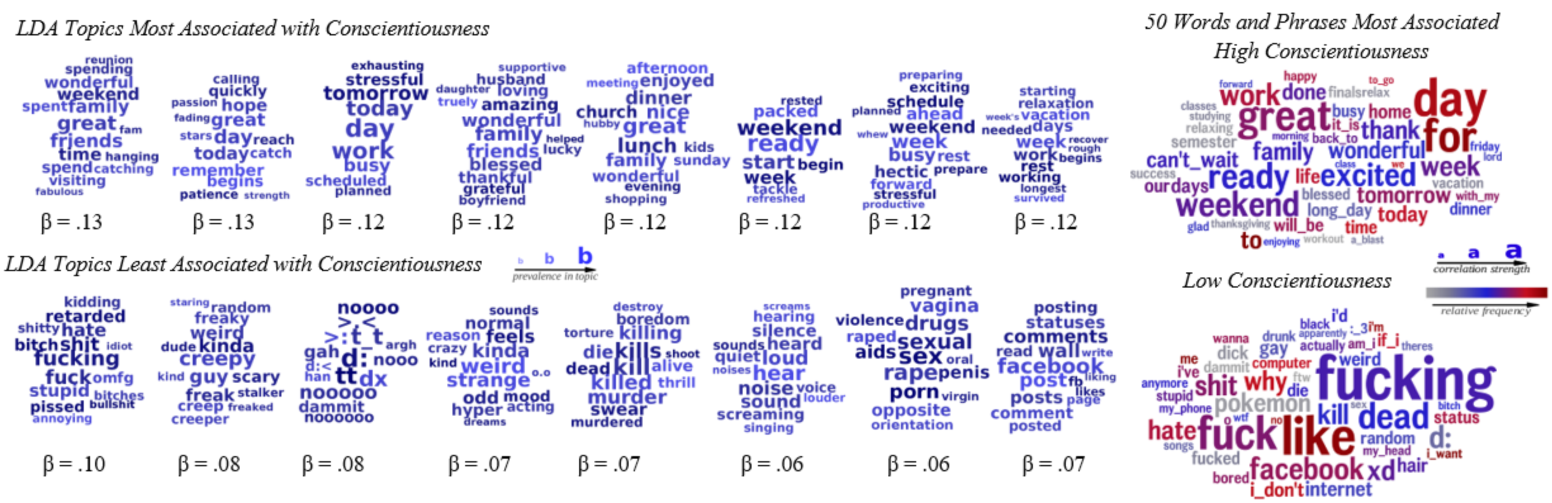


Table 6

Standardized Regression Coefficients between Extraversion and different Dictionaries across 65,896 Facebook users (controlled for Age and Gender)

	General Inquirer						DICTION		Linguistic Inquiry and Word Count (LIWC 2015)			
	Lasswell		Harvard IV		Stanford				LIWC (other)		LIWC (psych. processes)	
	Dictionary	β	Dictionary	β	Dictionary	β	Dictionary	β	Dictionary	β	Dictionary	β
More Ext.	Affect-Total		Pleasure	.12	Affiliation	.09	Optimism	.11	Emotional tone (m)	.18	Positive emotion	.16
	Affect-Other	.14	Children	.07	Positive	.08	+Satisfaction	.08	Clout (m)	.06	Drives	
	Affect-Domain	.12	Vary	.06	Strength	.04	+Praise	.05	Personal pronoun	.03	Affiliation	.12
	Affect-Gain	.10	Movement-Rise	.06	Active	.02	+Inspiration	.03	2nd person	.04	Reward	.09
	Affect-Participants	.04	Completion	.06			Insistence	.06	1st pers plural	.03	Netspeak	.11
	Positive-Affect	.07	Names	.06			Realism	.01	1st pers singular	.02	Social processes	.05
	Nations	.05	Emotion	.05			+Human Interest	.02			Friends	.09
	Power-Participants	.04	Travel	.05			+Temporal	.02			Family	.05
	(Ordinary)		Social Relation	.05			+Spatial	.02			Leisure	.07
	Wellbeing-Psych.	.04	Movement-Change	.05			Self-reference	.01			Future focus	.05
	Power-Cooperation	.04									Biological processes	.04
	Transaction-Gain	.04										
Less Ext.	Enlightenm.-Total	.06	Negation	.09	Weak	.05	Denial	.06	Negations	.06	Personal concern	
	Enlightenm.-Other	.08	Awareness	.08	Negative	.03	Hardship	.06	Auxiliary verbs	.06	Death	.10
	Enlightenm.-Ends	.08	Vice	.07	Understated	.03	Tenacity	.06	Personal pronouns		Work	.05
	Enlightenm.-Part.	.05	Abstract vocab.	.06			Ambivalence	.05	3rd pers plural	.06	Cognitive process	.09
	Denial	.06	Doctrine	.96			Activity		Impersonal pronouns	.05	Tentative	.09
	Uncertainty	.05	Comm. Tools	.06			-Cognition	.05	Common verbs	.05	Insight	.09
	Affect-Loss	.05	Change Finish	.05			+Communication	.03	Common adverbs	.05	Differentiation	.08
	Means	.05	Academic vocab.	.05			+Aggression	.03	Articles	.04	Causation	.07
	Negative-Value	.04	Pain	.05			Complexity	.04	Comparisons	.04	Risk	.08
	Negative-Affect	.04	Cardinal	.05			Familiarity	.04	Interrogatives	.04	Negative emotion	.07
							Exclusion	.04			Anxiety	.07

Note. All coefficients are significant at p < .001, corrected for multiple comparisons. (m) designates "master" categories that combine frequencies of multiple dictionaries.

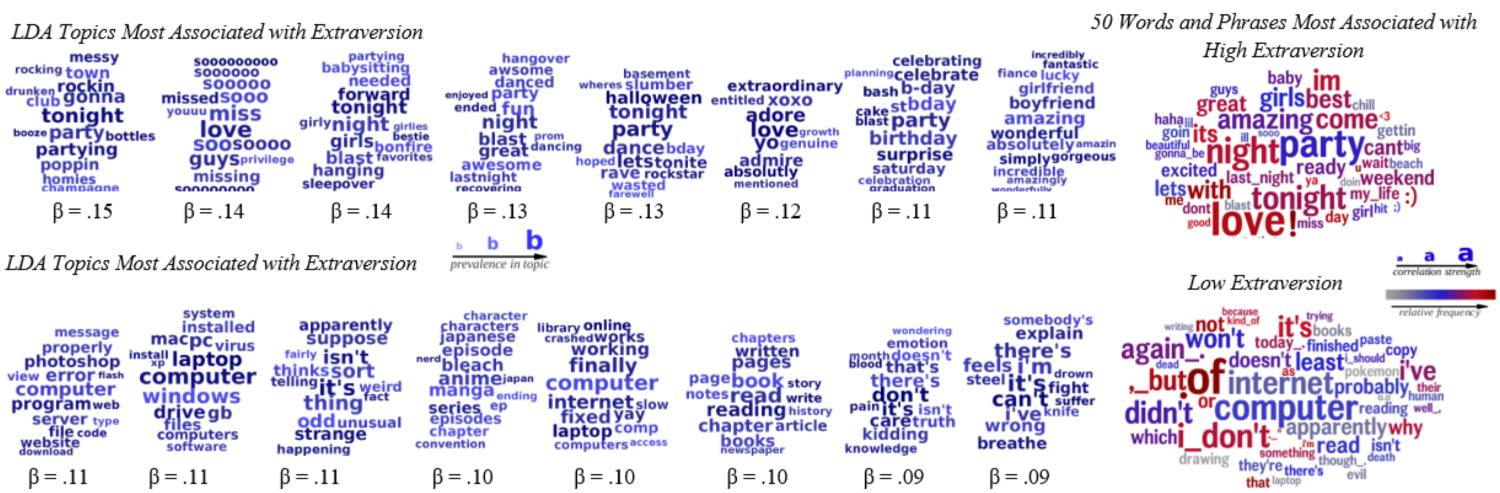


Table 7

Standardized Regression Coefficients between Neuroticism and different Dictionaries across 65,896 Facebook users (sample Age and Gender-stratified)

	General Inquirer						DICTION		Linguistic Inquiry and Word Count (LIWC 2015)			
	Lasswell		Harvard IV		Stanford				LIWC (other)		LIWC (psych. processes)	
	Dictionary	β	Dictionary	β	Dictionary	β	Dictionary	β	Dictionary	β	Dictionary	β
More Neu.	Negative-Affect	.10	Pain	.09	Weak	.07	Optimism		Negations	.07	Negative emotion	.15
	Affect-Loss	.06	Vice	.09	Negative	.07	-Hardship	.07	Common adverbs	.05	Anger	.11
	Wellbeing-Total	.03	Weak	.07	Passive	.04	-Blame	.05	Common verbs	.05	Sadness	.09
	Wellbeing-Loss	.05	Negation	.06	Hostile	.03	-Denial	.04	Personal pronouns	.03	Anxiety	.08
	Wellbeing-Phys.	.03	Need	.04	Understated	.02	Certainty		1st pers singular	.05	Death	.08
	Denial	.05	Self	.04			-Ambivalence	.05	3rd pers singular	.02	Cognitive process	.06
	Enlightenm.-Ends	.05	State Verb	.04			-Self-reference	.04	Auxiliary verbs	.04	Discrepancy	.07
	Negative-Value	.04	Awareness	.04			+Tenacity	.03	Conjunctions	.03	Tentative	.06
	Rectitude-Ethics	.03	Change-Finish	.04			Exclusion	.03			Biological processes	
	Enlightenm.-Other	.03	Disagreement	.03			Aggression	.02			Body	.06
							Present-Concern	.02			Sexual	.06
							Communication	.02				
Less Neu.	Affect-Other	.05	Pleasure	.07	Positive	.06	Optimism	.09	Emotional tone (m)	.17	Positive emotion	.10
	Nations	.04	Ritual	.07	Affiliation	.04	+Praise	.04	Clout (m)	.06	Drives	
	Power		Expressive	.06	Strength	.03	+Satisfaction	.03	Analytical thinking (m)	.06	Affiliation	.07
	Power-Coop.	.04	Places	.05	Power	.02	+Inspiration	.03	Personal pronouns		Reward	.07
	Power-Part.	.04	1st pers. plural	.05			Certainty	.03	1st pers plural	.05	Achievement	.06
	(Ordinary)		Names	.04			+Insistence	.05	Articles	.03	Personal concern	
	Power-Conflict	.03	Political	.04			+Collectives	.03			Leisure	.07
	Positive-Affect	.04	Land Places	.04			Temporal	.03			Religion	.05
	Respect-Lose	.03	Time-Broad	.04			Spatial	.02			Netspeak	.05
	Rectitude-Ends	.03	Travel	.04			Cooperation	.02			Relativity	.04
	Affect-Domain	.03									Time	.04
	Skill-Total	.03									Motion	.04

Note. All coefficients are significant at p < .001, corrected for multiple comparisons. (m) designates "master" categories that combine frequencies of multiple dictionaries.

LDA Topics Most Associated with Neuroticism

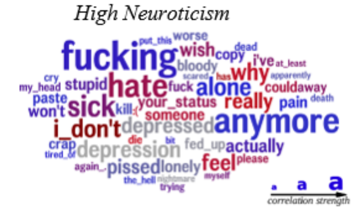
β = .09 β = .08 β = .08 β = .08 β = .08 β = .08 β = .07 β = .07

LDA Topics Most Associated with Neuroticism

β = .08 β = .08 β = .07 β = .07 β = .07 β = .07 β = .07 β = .07

50 Words and Phrases Most Associated with

High Neuroticism



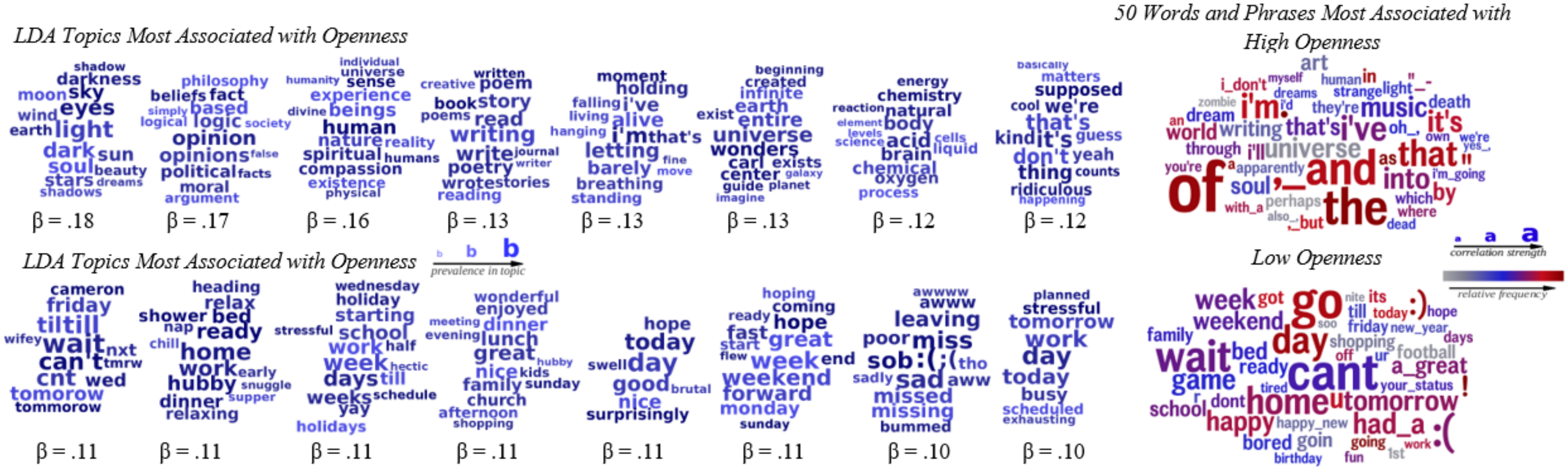
Low Neuroticism

Table 8

Standardized Regression Coefficients between Openness and different Dictionaries across 65,896 Facebook users (controlled for Age and Gender)

	General Inquirer						DICTION		Linguistic Inquiry and Word Count (LIWC 2015)			
	Lasswell		Harvard IV		Stanford				LIWC (other)		LIWC (psych. processes)	
	Dictionary	β	Dictionary	β	Dictionary	β	Dictionary	β	Dictionary	β	Dictionary	β
More Ope.	Skill-Aesthetic	.10	Awareness	.12	Understated	.06	Certainty		Articles	.15	Cognitive process	.09
	Enlightenm-Total	.07	Abstract vocab.	.10	Negative	.04	-Variety	.09	Total function words	.08	Insight	.12
	Enlightenm-Other	.09	Think	.09	Overstated	.04	+Tenacity	.07	Auxiliary verbs	.07	Causation	.07
	Enlightenm-Ends	.06	Doctrine	.08	Weak	.03	-Self-reference	.04	Comparisons	.06	Tentative	.07
	Arenas	.08	Quality	.08	Passive	.02	-Ambivalence	.04	Impersonal pronouns	.06	Death	.12
	Form	.07	Perceive	.07			Complexity	.11	Conjunctions	.06	Perceptual process	.12
	Power-Authority	.06	Nature-Process	.07			Familiarity	.07	Prepositions	.05	Hear	.08
	Power-Participants	.05	Independent Adj.	.07			Cognition	.06	1st pers singular	.04	See	.07
	(Ordinary)		Negation	.07			Centrality	.06	Interrogatives	.04	Anxiety	.08
	Wealth-Total	.05	Evaluation2	.07			Exclusion	.06	Quantifiers	.04	Space	.05
	Wealth-Other	.06					Communication	.04				
Less Ope.	Affect		Kinship	.10	Submit	.07	Certainty	.02	Emotional tone (m)	.08	Netspeak	.14
	Affect-Other	.08	Persistence	.10	Affiliation	.04	+Insistence	.13	Clout (m)	.05	Family	.13
	Affect-Participants	.05	Pleasure	.08			+Collectives	.02	2nd person	.02	Affective process	.10
	Affect-Domain	.05	Time (Broad)	.07			Realism	.02			Positive emotion	.11
	Power-Cooperation	.07	Movement-	.07			+Temporal	.07			Drives	
	Well-being Total	.04	Change (Stay)				+Human Interest	.02			Reward	.11
	Wellbeing-Psych.	.07	Social	.06			Optimism	.04			Affiliation	.08
	Wellbeing-Participants	.04	Ritual	.06			+Satisfaction	.04			Future focus	.09
	Respect-Lose	.06	Try	.05			+Praise	.03			Home	.08
	Positive-Affect	.05	Vary	.05			Motion	.04			Relativity	.05
	Nations	.04	Travel	.05							Time	.10

Note. All coefficients are significant at p < .001, corrected for multiple comparisons. (m) designates "master" categories that combine frequencies of multiple dictionaries.



Personality. Tables 4-8 show the dictionaries, word and phrases, and LDA topics most associated with the users' personality scores across the Big Five personality dimensions. Associations between personality and language variables were markedly weaker than those for age and gender (β ~ .20 for the most associated language features, versus β ~ .30 for age and gender). The most consistent and often strongest associations were with positive and negative emotion dictionaries.

Agreeableness. As shown in Table 4, Agreeableness demonstrated the strongest associations with positive emotion and optimism. It was weakly associated with greater use of first person plural pronouns (GI-Harvard-IV: *first person plural* and LIWC: *first person plural*, β s = .06). It was also weakly associated with dictionaries reflecting affiliation (GI-Stanford: *affiliation*, β = .06; LIWC: *affiliation*, β = .09,), aligned with other studies (Jensen-Campbell, Knack, & Gomez, 2010). Low agreeableness was dominated by swear words.

Conscientiousness. As shown in Table 5, Conscientiousness was positively associated with references to work and economic concerns (GI-Harvard-IV: e*conomic*, β = .07; GI-Lasswell t*ransaction-gain,* β = .07; LIWC: *work,* β = .11). While the words and phrases include words reflecting work, they also include positive emotion, family, and a sense of relaxing from work.

Extraversion. As shown in Table 6, like Agreeableness, Extraversion was weakly associated with greater use of positive emotion and affiliative dictionaries.

Neuroticism**.** As shown in Table 7, across the different dictionaries, Neuroticism was most strongly associated with expressions of positive (inversely) and negative emotions, as might be expected. The topic, words, and phrases further results help to specify processes underlying these findings. Topics reflect somatic concerns (*feeling, tired, sick,)*, hostility and cursing (*fuck, asshole*), but also exhaustion and over-arousal (*stressed*, *frustrated*, *annoyed)* and low mood and self-esteem, reminiscent of dysphoria and depression (*lonely, depressed*, *hopeless).* Beyond positive emotions (*awesome, amazing, exciting),* the language most associated with emotional stability includes *weekends* as well as sports (*workout, football*, *team, game)* and religious practices and affiliation (*blessed*, *lord*, *Jesus*). Weekends and religion are also captured by the LIWC *leisure* (r = .07) and *religion* (*r* = .05) dictionaries.

Openness. As Table 8 suggests, Openness was positively associated with cognition-related dictionaries (GI-Harvard-IV: *awareness*, β = .12, *abstract vocabulary*, β = .10; GI-Lasswell: *enlightenment-total*, β = .07; LIWC: *insight*, β = .12), reflecting intellect and insight. The DLA words and phrases reflect greater lofty, abstract, and transcendental language (*soul*, *universe, dream*). Low openness was related to dictionaries reflecting time orientation (GI-Harvard-IV: *time-broad*, β = .07; DICTION: *temporal*, β = .07; LIWC: *time*, β = .10), family (GI-Harvard-IV: *kinship*, β = .10; LIWC: *family*, β = .13), and home (LIWC *home,* β = .08). These concepts are similarly mirrored in the DLA results (*home, today*, *tomorrow, week, weekend)*.

Predictive Power

To quantitatively gauge how much gender, age, and personality variance in the language domain is captured by the different sets of language variables, we examined the cross-validated prediction performances of prediction models that used the different sets of language variables (GI, DICTION, LIWC, and 2,000 LDA topics) as features, as well as a more sophisticated models that combined topics, words, and phrases as features (see

Park et al., 2014 and Sap et al., 2014 for details on the method).

As shown in Table 9, Diction's 36 language categories captured markedly less information about personality (average r = .18) than LIWC (r= .27) and GI (r = .28), suggesting that their dictionaries capture similar amounts of personality-relevant information. Given the fact that LIWC has only about a third the dictionary categories of GI, it appears more parsimonious while equally exhaustive. The LDA-topic-based prediction performances were about 20% higher ($\Delta r \sim .06$) than those achieved by GI and LIWC, and 10% lower ($\Delta r \sim .04$) than sophisticated prediction models using many more language features. The adjusted R^2 for LIWC, GI, and the LDA topics was evenly matched (R^2 = .07, .08, .09, respectively). Altogether, the 2,000 LDA topics captured the most personality-related variance in language.

Table 9

Cross-validated prediction performances of Prediction Models Using the Dictionaries of the Different Software Programs.

	Diction	LIWC 2015	Gen. Inquirer	LDA Topics	LDA Topics, Words, Phrases
Number of predictors	36	73	182	2,000	(varies)
Age (r)	.43 (.42, .43)	.65 (.65, .66)	.68 (.68, .69)	.79 (.79, .80)	.83 [a]
Gender (accuracy)	.70 (.69, .70)	.78 (.78, .79)	.81 (.81, .82)	.88 (.88, .88)	.92 [a]
Personality					
Agreeableness (r)	.16 (.15, .16)	.26 (.25, .26)	.24 (.23, .25)	.29 (.29, .30)	.35 [b]
Extraversion (r)	.16 (.16, .17)	.27 (.27, .28)	.29 (.28, .29)	.36 (.35, .36)	.42 [b]
Conscientiousness (r)	.21 (.20, .22)	.30 (.29, .30)	.30 (.29, .31)	.34 (.34, .35)	.37 [b]
Neuroticism (r)	.14 (.13, .14)	.24 (.23, .24)	.26 (.25, .27)	.31 (.30, .32)	.35 [b]
Openness (r)	.24 (.23, .25)	.30 (.30, .31)	.32 (.31, .33)	.39 (.38, .40)	.43 [b]
Average Pers. Correlation	.18	.27	.28	.34	.38
Average Pers. Adj. R2	.03	.07	.08	.09	

Note: For continuous outcomes, prediction performance is given by the Pearson correlation between the predicted values and the actual values. For gender, performance is given by classification accuracy of a penalized logistic regression model. The column on the right gives the state-of-the-art performances for comparison. Parentheses indicate 95% confidence intervals ([a]Sap et al., 2014, [b]Park et al., 2014). LIWC 2015 predictions were based on the dictionaries provided with LIWC 2015, applied to the word frequency counts through our Python code base. The LIWC software extracts additional language variables, including meta-features and composite variables, which when included in a prediction model produced the same average prediction performances across personality traits as the Python-derived frequencies.

Power Analyses

Figure 4 illustrates the average number of features from the different language sets significantly associated with age and gender (top) or personality (bottom) across different sample sizes of Facebook users with at least 1,000 words each. As a rough guide, the exploratory language analyses produced findings of theoretical nuance with about 10 significantly associated LIWC dictionaries, 100 out of 2,000 LDA topics, or 200 out of 11,894 words and phrases. Table 10 provides estimates on sample sizes needed (with 1,000 words each) to reach this number of significant features for gender, age, and personality. For personality, across 1,000 users 10 LIWC dictionaries and 100 LDA topics were significantly associated, while 200 significant words and phrases required the power of 3,000 users.

There was also substantial variance between the different Big Five factors; for example, 500 users sufficed for 10 significantly associated LIWC dictionaries for Conscientiousness, while 1,500 users were needed for Neuroticism. As larger regression coefficients were observed for age and gender than for personality, more significant associations can be observed in smaller samples.

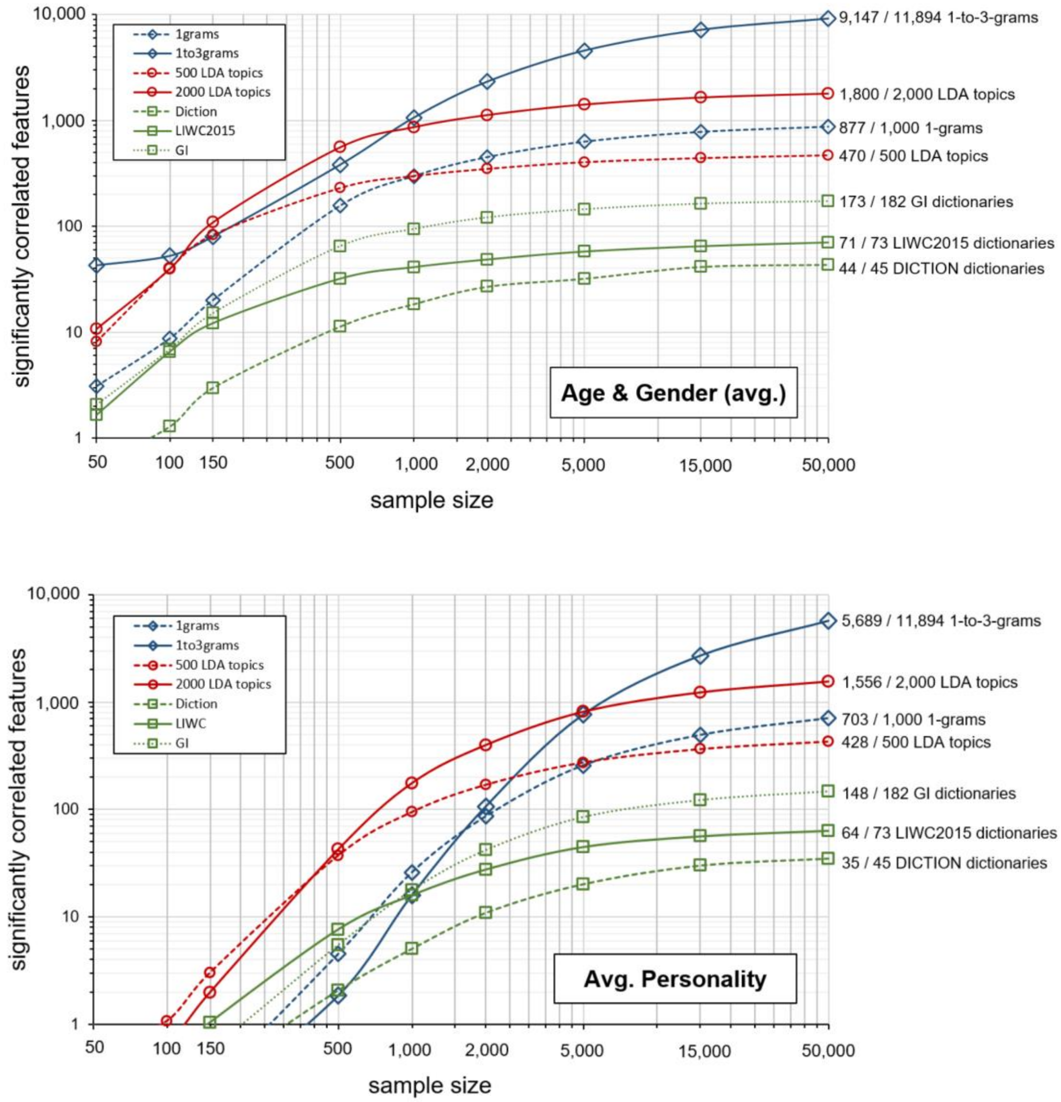


Figure 4. Average number of language features significantly associated with age and gender (top) and Big Five personality (bottom) as a function of the number of included users (sample size) for different feature sets (age associations controlled for gender and vice versa, personality regressions controlled for age and gender). For sample sizes of 50 to 150, the significantly associated features shown are the average of 100 random draws from the overall sample (*N* = 65,986); sample sizes of 500, 1,000, 5,000, 15,000, and 50,000 are based on 50, 20, five, three, and one random draws, respectively.

Table 10.
Minimal Sample needed for Exploratory Language Analyses

Thresholds of significant correlates:	Demographics		Big Five Personality					
	Gender	Age	Agr.	Con	Ext.	Neur.	Ope.	**(avg.)**
10 (out of 73) LIWC dictionaries	200	150	800	400	800	1,100	550	**750**
100 (out of 2,000) LDA topics	250	150	1,100	550	800	1,800	550	**1,000**
200 (out of 11,894) 1-to-3 grams	650	200	3,650	1,850	2,600	4,750	2,100	**3,000**

Note. Sample sizes (N) needed of Facebook users to observe 10 significantly associated LIWC dictionaries (out of 73), 100 LDA topics (out of 2,000), or 200 1-to-3 grams (out of 11,894) for gender, age, and personality (using all of the users' Facebook posts). Significance threshold of alpha = .05 was Benjamini-Hochberg corrected for multiple comparisons.

Choosing the Number of Topics to Extract

In the topic modeling process, the user may choose the numbers of topics to extract, adjusting specificity. Topics disambiguate different word senses, and a larger number of topics can provide more fine-grained context distinctions, but can also increase the number of repetitive topics. Table 11 shows the topics that have the word *play* among their top 10 words, across topic sets of 50, 500 and 2,000, modeled over the same 5 million statuses. While 50 topics failed to distinguish *ball play*, *musical play*, and *videogame play*, 500 topics successfully distinguished these contexts. The 2,000 topics distinguished different kinds of video games (i.e., military first-person shooters *Call of Duty: Black Ops*, real-time strategy *Starcraft*, and the action-adventure game *Assassin's Creed*]. Finally, Figure 5 illustrates prediction accuracies using 50, 500, and 2,000 topics, modeled across varying numbers of Facebook statuses, when applied to the language of all 65,896 users and used to their personality. The prediction models based on 500 or 2,000 topics were comparable, and outperformed those built over 50 topics.

Table 11

Topics Mentioning Play for Sets of Topics of Different Sizes.

Top. Set	Occ.	Top 10 words comprising each topic
50	1	game, play, win, playing, football, team, won, games, beat, lets
500	5	guitar, play, playing, music, piano, band, bass, hero, practice, played
		game, football, play, soccer, basketball, playing, games, team, practice, baseball
		play, playing, game, ball, games, played, golf, tennis, poker, cards
		play, playing, game, games, xbox, halo, wii, video, mario, 360
		place, chuck, find, meet, play, birth, norris, interesting, babies, profile
2000	9	play, guitar, learn, piano, learning, playing, learned, lessons, songs, rules
		play, game, let's, role, sims, rules, chess, basketball, plays, poker
		play, playing, tennis, cards, wii, played, poker, ball, basketball, pool
		soccer, football, game, play, team, basketball, playing, ball, practice, field
		black, cod, ops, playing, play, mw2, modern, warfare, ps3, online
		play, playing, starcraft, warcraft, sims, ii, beta, online, nerds, nerd
		xbox, 360, play, ps3, playing, games, creed, assassin's, playstation, assassins
		words, comment, note, play, wake, jail, copy, paste, sport, fair
		games, play, playing, game, video, played, card, board, begin, playin

Note. Top ten words for topics that included “play” among their top 10 words for sets of 50, 500, and 2,000 topics modeled over the same 5 million Facebook statuses. Words suggesting playing music are highlighted in green, ball sports in blue, and videogames in yellow.

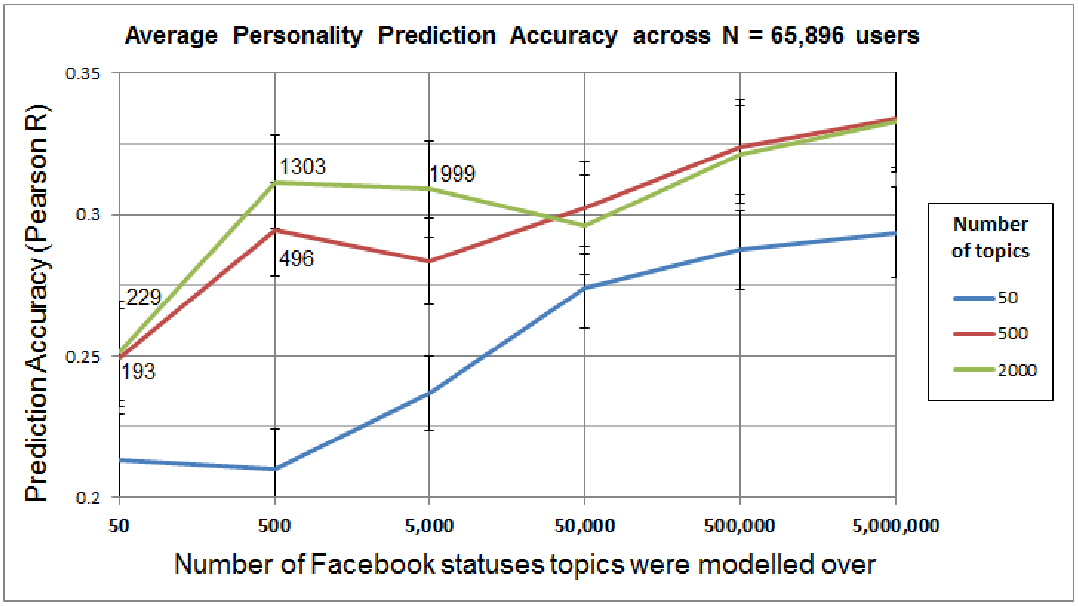


Figure 5. Prediction accuracies (across 65,896 users and 12.7 million Facebook statuses) obtained using 50, 500, and 2,000 topics, modeled across 50 to 5 million Facebook statuses. Cross-validated ridge-regression prediction accuracies were averaged across the five personality traits; error bars give the standard error of the average. When the number of topics to be modeled was close to or exceeded the number of statuses to be modeled over, the MALLET package created fewer topics; in those case the actual number of topics modeled is noted.

Discussion

This review quantitatively compared three closed-vocabulary sets of dictionaries (provided by the General Inquirer, DICTION, and Linguistic Inquiry and Word Count) with two open-vocabulary methods (Latent Dirichlet Allocation and Differential Language Analysis) across 13 million Facebook status updates from 65,000 users. GI, DICTION, and LIWC dictionaries associations were larger for age and gender than for Big Five personality. Open-vocabulary results were congruent with but conceptually more specific than dictionary associations. Cross-validated machine learning prediction models indicated that the 2,000 LDA topics provided superior predictive power, and thus captured more demographic- and personality-related variance in language.

The language results corroborate and expand previous studies on the association of language with age (e.g., Pennebaker & Stone, 2003; Schwartz et al., 2013b), gender (e.g., Newman, Groom, Handelman, & Pennebaker, 2008; Schwartz et al., 2013b), and personality (Kern et al., 2014a; Schwartz et al., 2013b; Yarkoni, 2010). GI, DICTION, and LIWC overlap in their coverage of pronouns and concepts, including positive and negative emotion, complex language suggestive of higher cognition, economic and fiscal concerns, and social and family relationships. The dictionaries that distinguished emotional valence were among the most associated dictionaries with female gender, older age, higher levels of Agreeableness, Conscientiousness, Extraversion, and lower levels of Neuroticism. Prediction models based on GI and LIWC dictionaries reached similar prediction performances, and out-predicted DICTION.

Similar to previous work (Iacobelli, Gill, Nowson, & Oberlander, 2011; Schwartz et al., 2013b), the open-vocabulary prediction models based on 2,000 LDA topics significantly outperformed dictionary-based prediction models, suggesting that the larger number of open-vocabulary features capture more of the personality-related variance in the language data. Modeling and extracting a greater number of topics has clear advantages (more specificity) and only a limited disadvantage that can be handled algorithmically (more duplicate topics, which can be filtered).

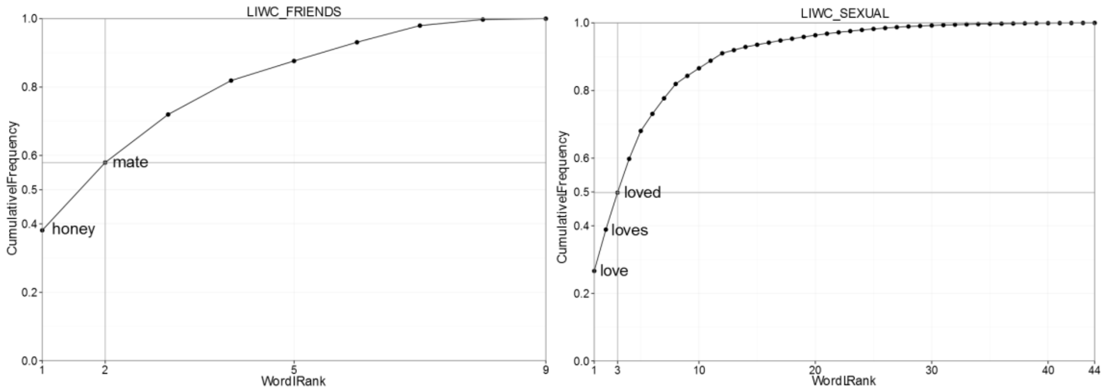


Figure 6. Cumulative frequency distributions of the *LIWC 2007 friends* (left) and *LIWC 2007 sexual (right)* dictionaries. 50% of the dictionary counts are due to four words or less in both cases, and the leading words in the dictionaries are word-sense-ambiguous.

Dictionary Based Text Analysis: Sources of Error

Dictionary-based word count programs have become the default method to analyze textual data in psychology. These programs have provided numerous insights. However, the programs also bring a number of sources of error.

A few words drive a dictionary. As others have noted (Alderson, 2007; Chung & Pennebaker, 2007; Pennebaker, 2011), a few words often make up the majority of occurrences in the English language. Most words occur rarely and the majority of occurrences in a dictionary can often be attributed to a small number of words. In the current study, 96 words made up more than 50% of word occurrences (Figure 1). As an example, about dictionary frequencies depending on a very small number of words, in the previous and most-cited version of LIWC (2007), two words (*honey*, *mate*) accounted for more than 50% of the occurrences of the *friends* dictionary. Three words (*love*, *loves*, *loved*) accounted for 49.8% of the occurrences of the LIWC 2007 *sexual* dictionary (see Figure 6; the LIWC 2015 *friends* and *sexual* dictionaries no longer include these words).

When these highly frequent words are ambiguous--as they are here--and have primary word senses that do not match the concept intended by the creator of the dictionary, the dictionary results can be misleading.

Other sources of error. Beyond word-sense ambiguities, all methods used here use a *bag of words* approach. Words are counted regardless of their context, including negation or irony. In previous work (Schwartz et al., 2013c), raters examined 100 Facebook statuses that contained words from the LIWC *positive* and *negative emotion* dictionaries, but were rated as Type I errors (i.e., false positives). Table 12 reports the relative frequencies of sources of errors. About 50% of such false positives were due to lexical ambiguities (word sense and part of speech ambiguities), 21% was due to negation, and 30% was due to other sources. Type II errors (false negatives) occur when dictionaries fail to identify instances of the expression of the psychological construct they are intended to measure, and are more likely to reduce observed effect sizes (low "recall"). Type II errors can often be remedied with larger sample sizes. To estimate the false positive error rate of dictionaries, human raters should validate dictionaries for a language corpus by rating if the occurrence of dictionary words correctly mark the dictionary concept intended, particularly if the dictionary findings are critical to the argument being made.

Table 12
Sources of Error in LIWC Positive and Negative Emotion Dictionaries.

Category	Source of Error	% of Cases	Description	Examples
Lexical Ambiguity	Wrong part-of-speech	14%	Not a valid signal because it is the wrong POS	So everyone should come to the **play** tomorrow…
	Wrong word sense	37%	Not a valid signal because it is the wrong word sense (includes metaphorical senses)	Does anyone know what type of file I need to convert youtube videos to **play** on PS3??
Signal Negation	Strict negation	15%	Within the scope of a negation, where there is a clear negative quantifier	… all work no **play** :(
	Desiring	6%	Within the scope of a desire / wishing for something	I sure wish I had about 50 hours a day to **play** cod
Other	Stem Issue	28%	Clearly not intended to be matched with the given stem	**Numb*** for *NEGEMO* matching *number*
	Other		Any other issue or difficult to classify	

Note. Distribution of errors across 100 Facebook statuses in which words contained in the positive and negative emotions dictionaries were rated as not expressing those emotions. Adapted from Schwartz, et al., 2013b, Table 3 & 5.

Recommendations for Researchers

Our quantitative review suggests a series of recommendations to consider when analyzing text data.

Choosing an approach. Dictionary based word-count programs have been instrumental in adding text analysis to the toolbox of research psychologists. Open-vocabulary data-driven methods like LDA topic-modeling have been developed in Natural Language Processing that provide a valuable complement. Given both dictionary-based and open-vocabulary methods, which method should one use? If possible, *both*.

Dictionary-based text analysis has a number of properties that make it desirable: (a) as the dictionaries are the same across studies, results are comparable and (b) a set of dictionaries yields a relatively parsimonious quantitative representation of language content. Validated dictionaries can be suitable for *testing specific hypotheses*. But dictionary based approaches also have numerous sources of potential errors, like the disproportionate impact of highly frequent but ambiguous words, which can be addressed through dictionary validation.

Open-language approaches are desirable because they (a) yield more specific language findings that are suitable for the generation of specific hypotheses (e.g., specific emotions); (b) capture more construct-related variance in the language (i.e., have higher predictive power); and (c) they can help unpack dictionary-based findings. Open-vocabulary results can be shortlisted, filtered for uninformative duplicates, and visualized for inspection as a list or word cloud, yielding interpretable and intuitive summaries of the language most distinguishing of a trait.

However, word, phrase and topic extraction can be harder to implement and requires more expertise. In addition, many function word categories (like pronouns) cannot suitably be captured through topic modeling; their omnipresence in the language across different contexts would add them to most topics. Thus, such highly frequent words are routinely excluded from the analysis when topics are modeled (as they were in this analysis). Function word dictionaries offer a simple and parsimonious way to keep them as units of analysis. Further, even when conducting open-vocabulary analyses, examining the associations of a given trait with a set of dictionaries allows the researcher to quickly get a sense of the language correlates of a given trait, before examining a potentially large number of topic correlations in more detail. In this way, dictionary-based correlations can help the researcher see the broad patterns behind the specific word, phrase and topic correlations, providing a first step for triangulating on the full pattern of results. In our own work we have found the combined used of these methods invaluable for seeing the whole story in the language data.

Sample size considerations. Perhaps surprisingly, for exploratory language analyses, even when correcting significance thresholds for multiple comparisons, an analysis with 2,000 LDA topics does not require a substantially larger sample than using 73 LIWC 2015 dictionaries (~200 Facebook users for age and gender, 1,000 vs. 750 users for Big Five personality; see table 10). Previous findings suggest that to the order of 500-1,000 words are needed per user for dependable language estimates (Kern et al., 2016).

For Differential Language Analyses with words and phrases (1-to-3 grams), substantially larger samples are need to explore the differences in language use across gender (~650) and personality (~3,000 users), while appropriately controlling for multiple comparisons.

Dictionary considerations. Most words only negligibly contribute to the overall dictionary word-count. When the few highly frequent words predominantly occur in a text sample in a different word sense than was intended by the dictionary creator, interpretations based on the dictionary frequencies can be invalid. Thus, dictionaries should be validated for a given language sample, particularly when the validity of a given dictionary is essential for the analytic strategy.

To validate a dictionary in a given study, one or more human raters should examine instances in which a language unit of analysis (like a sentence, Tweet, or Facebook status) contains the words in a given dictionary, and rate as to whether the language unit of analysis expresses the concept intended by the dictionary. The dictionary accuracies should be reported in the methods or results. For example, Schwartz et al.

(2013) found LIWC's (2007) popular positive and negative emotion dictionaries to mark expression of positive and negative emotion correctly with about 70% accuracy in Facebook statuses. Eichstaedt et al. (2015) found that the LIWC *anger* and *anxiety* dictionaries had accuracies of 60% and 55%, respectively (across 100 Tweets).

Given that dictionaries are often determined by a few highly frequent words, and about 50% of the false positives are due to lexical ambiguities, determining as to whether a given dictionary's most frequent word's most frequent word-sense captures the dictionary concept may be a good place to start (see table S1 in Appendix A for such statistics for LIWC 2015). But whenever a dictionary is applied to new language contexts other than those for which it was designed, Grimmer and Stewart's (2013) advice should be followed: "Validate, validate, validate" (p. 3).

Topic model considerations. In 2003, Pennebaker, Mehl and Niederhoffer wrote:

> Although not emphasized in this article, word count strategies are generally based on experimenter-defined word categories. These categories are based on people's beliefs about what words represent. Hence, they are ultimately subjective and culture bound. Content-based dictionaries that are aimed at revealing what people are saying have not yielded particularly impressive results owing in large part to the almost infinite number of topics people may be dealing with. With the rapidly developing field of artificial intelligence, the most promising content or theme-based approaches to text analysis involve word pattern analyses such as LSA. These purely inductive strategies provide a powerful way to decode more technical or obscure linguistic topics. For

> researchers interested in learning what people say—as opposed to how they say it—we recommend this new analytic approach (p. 571)

LDA topic modelling was developed in the same year in which the above passage was written (Blei, Ng & Jordan, 2003) and has succeeded LSA as the most popular analytic (Landauer, Foltz, & Laham, 1998) strategy for data-driven text mining. It yields semantically coherent topics (clusters of words) based on patterns of word co-occurrence that implicitly disambiguate the different word senses of ambiguous words (for examples, see Table 11). Topics have the advantage of keeping individual words with their context. A cluster of words in a topic around a consistent theme can be a more dependable unit of analysis than single word associations, or dictionaries that are dominated by ambiguous, highly frequent words. Creating topics based on a given language corpus is also an efficient way of summarizing the themes mentioned in the corpus.

Generally, the larger the corpus, the more coherent and fine-grained the resulting topic models are. All things being equal, our analysis suggests that one ought to err on the side of modeling more (500+) rather than fewer topics on a given corpus.

Notably, it is not necessary to develop the topics on the same language dataset to which they are applied. This creates the possibility of creating topic models on a larger language sample (and thus contain more content to inform the modeling process), and then applying the topics to a smaller study sample, much like the dictionary approach, but driven from the data rather than from theory. Using the same set of topics across multiple studies and datasets can also allow researchers to compare topic results across datasets (for example, the 2,000 LDA topics used in this study were previously used to analyze county-level Twitter language (Eichstaedt et al., 2015; Schwartz et al., 2013a).

Resources and tools. Part of LIWC's success story has been the ease of use of the program. While many packages exist to perform topic modeling, none of them currently is as easy to use as LIWC. To help make these methods more accessible, we have created an online tool with which users can extract the 500 and 2,000 topics used in this study from their text samples which may be uploaded in the LIWC input format. We are also releasing the 500 and 2,000 topics in the form of weighted dictionaries that can be used as part of other text analysis programs[3], as well as the General Inquirer dictionaries that capture as much trait-related variance as LIWC, but are free for non-commercial use (for all resources, see http://lexhub.org/tools and http://wwbp.org/data.html). Differential Language Analysis can be carried out using the open-source Python code base we have released for non-commercial purposes (see http://dlatk.wwbp.org).

Limitations

While this review compares three dictionary approaches and two open-vocabulary approaches, it does not address the ways in which supervised machine learning methods might augment or even replace annotation by humans (for a thoughtful review of this point, see Grimmer & Stewart, 2013), or how dictionaries could be improved using data-driven approaches (e.g., Sap et al., 2014, Schwartz et al. 2013). We do not discuss the many other emerging algorithms to create topic models that take author attributes into account, or cluster words based on embeddings, such as Word2Vec. We also omitted a

[3] Unfortunately LIWC2015 does not support weighted dictionaries.

discussion of how dimensionality reduction techniques can be combined (for example, multi-level LDA, or a combination of exploratory factor analysis and LDA topic modeling) to create a more parsimonious representation of the language space.

Conclusion

Text analysis in psychology is at a methodological juncture: the literature thus far has relied almost entirely on closed-vocabulary programs with predetermined dictionaries, yet recent innovations promise to complement or even in-part replace these traditional programs with data-driven methods.

DICTION's method of combining multiple dictionaries into master variables is not recommended, as the results can be impossible to interpret. The General Inquirer was ahead of its time and provides dictionaries on par in quality and coverage (but not parsimony) with LIWC, and its dictionaries are free for non-commercial use. Many (but not all) dictionaries provide reliable measures of their intended constructs. But because of the Zipfian distribution of language and lexical ambiguities, no dictionary should be taken at face value--especially when it used in a different language domain than the one for which it was intended. Dictionaries of function words (like pronouns) are powerful markers of underlying cognitive and attentional psychological processes, and together with positive and negative emotion dictionaries are often among the most distinguishing markers for personality and demographic traits. Topic models like LDA--either modeled on the same corpus or imported from a larger one--produce more fine-grained, contextually embedded, and more transparent units of analysis than do dictionaries.

The largest datasets of our digital era are textual in nature. Learning how to process text at scale will be the price to pay to access the largest longitudinal, cross-sectional, and cross-cultural study in human history. Both closed and open-vocabulary approaches are needed to allow psychologists to test their hypotheses, and to discover new ones.

The previous chapter reviewed traditional dictionary-based methods of text analysis, and compared them to modern open-vocabulary approaches borrowed from Natural Language Processing in computer science. The following chapters turn to the prediction and characterization of health through social media sources using the methods discussed in the first chapter. The second chapter reviews the recent literature on mental health prediction across the three major sources of text on the web: Facebook, Twitter and web forums. As most of the studies discussed in this chapter are published in computer science venues rather than psychology journals, they tend to focus on the relative performance of prediction algorithms rather than trying to characterize the language correlates of mental illness in-depth. The study presented in the third chapter will use Facebook to predict the depression status of patients, and use the previously introduced open and closed-vocabulary methods to provide a more fine-grained analysis of the specific language markers predictive of depression in the study sample.

CHAPTER 2

DETECTING MENTAL ILLNESS THROUGH SOCIAL MEDIA: A REVIEW

A growing number of studies examine mental health in social media contexts, linking social media use and behavioral patterns with stress, anxiety, depression, suicidality, and other mental illnesses. The greatest number of studies focus on depression. Most studies either examine how the use of social media sites correlates with mental illness in users (Seabrook, Kern, & Rickard, 2016) or attempt to detect mental illnesses and symptoms from social media – the latter form the focus of this review.

Although diagnoses of depression and other mental illnesses have improved over the past two decades, they remain under-diagnosed detected, in part due to stigmas around seeking help for mental health concerns. Automated analyses of social media potentially provide potential early detection systems, and integrated with treatment. For example, if an automated process detects elevated depression scores, that user could be targeted for a more thorough assessment, and provided with further resources, support, and treatment.

Assessment

Methods used in these studies for identifying users with a mental illness included either recruiting participants to fill out one or more depression inventories, searching public Tweets for individuals who claim to have been diagnosed with depression, studying the language used in mental illness forums, or manual coding of social media posts for relevant mentions of mental illness (see Fig. 1). No study utilized clinician judgment or the "gold standard" for diagnosis, a semi-structured interview delivered by a clinician (American Psychiatric Association, 2013). As such, it should be noted that the studies reviewed here are based on mental illness screenings only, not diagnoses.

Prediction

Each study aimed to predict mental illness using social media, but they differed in how the prediction tasks were set up and evaluated. Prediction performances are generally evaluated in a cross-validation framework, in which prediction models are trained and tested on separate parts of the data (see Table 1 for prediction performances). Some studies established two balanced classes, with an equal number of "depressed" as "non-depressed" users, while others used mental illness base rates closer to their estimated distribution in the population (U.S. prevalence rates below 10%; National Institute of Mental Health, 2015). In the former it is easier to achieve high performance, but this approach runs the risk of lacking ecological validity. The choice of performance metric matters: in a sample with 20% depressed users, a simple decision rule of judging all users healthy would achieve 80% *accuracy*. In contrast, *Areas Under the ROC Curve* (AUCs) incorporate a comparison of false positive to false negatives rates and do not depend on class balance, and are thus in principle more comparable across studies and prediction tasks (highlighted in green in Table 1).

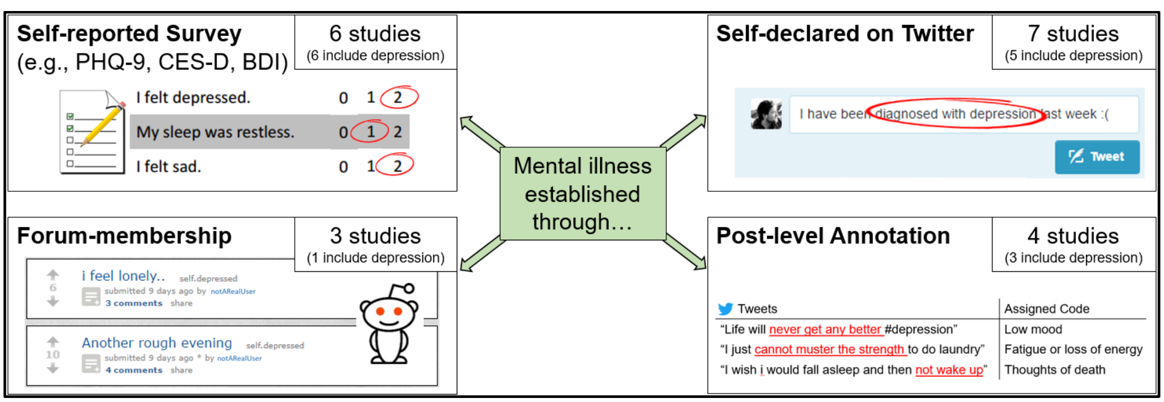


Figure 1. Criteria used by different sets of studies to establish mental illness status. Numbers of studies selected in this review are given, and only counted as including

depression if they did so as a separate condition.

Prediction of Survey Responses

Six studies relied on self-reported measures. The most cited study used Twitter activity to examine network and language data preceding a recent episode of depression, which was determined based on the self-reported presence and date of recent episodes of depression, and scores on the CES-D and BDI (De Choudhury, Gamon, Counts, & Horvitz, 2013). This study revealed differences in posting activity between depressed and non-depressed users including different diurnal cycles, more negative emotion, less social interaction, more self focus, and increased posting about depression terms throughout the year preceding depression onset. A similar prediction model was applied to the Tweets of US states and 20 US cities to derive population-level depression estimates (De Choudhury, Counts, & Horvitz, 2013)**.**

In Reece et al., (2016), user depression and post-traumatic stress-disorder (PTSD) status were predicted with comparably high AUC scores (.87/.89) from text and Twitter metadata preceding a reported first episode. Data were aggregated to weeks, which somewhat outperform aggregation to days, and could be modelled as longitudinal trajectories of activity patterns that differentiated healthy from mentally-ill users. In Tsugawa et al., (2015), depression prediction was reproduced in a Japanese sample, finding that prediction performance did not improve with additional data beyond 500 to 1,000 tweets from a person collected in the 2 to 4 months preceding the administration of the CES-D**.**

This work can be extended to Facebook posts. In De Choudury, Counts, Horvitz, & Hoff (2014)**,** self and survey-reported post-partum depression (PPD) were predicted, finding that 35.5% of the within-sample variance in PPD status could be accounted for by demographics, pre-partum Facebook activity, and content of posts. In Schwartz et al., (2014), questions from a personality survey were used to determine users' continuous depression scores across a much larger sample (N = 28,749), detecting seasonal fluctuations.

Prediction of Self-Declared Mental Health Status

Seven studies relied on users who publicly shared information about their mental illness diagnosis on Twitter. Computational Linguistics and Clinical Psychology (CLPsych) workshop was started in 2014 to foster cooperation between clinical psychologists and computer scientists. Datasets were made available and "shared tasks" designed to explore and evaluate different solutions to a shared problem. In the 2015 workshop, participants were asked to predict if a user had PTSD or depression based on self-declared diagnoses (PTSD = 246, depression = 327, with the same number of age- and gender-matched controls) (Coopersmith, Dredze, Harman, Hollingshead, & Mitchell, 2015). Participating teams built topics by considering all tweets from a given week as one document to build topic models (Resnik et al., 2015), grouped binary unigram vectors to apply Differential Language Analysis (Preotiuc-Pietro et al., 2015)**,** considered sequences of characters (Coopersmith, Dredze, Harman, Hollingshead, & Mitchell, 2015), and applied a rule-based approach to examine raw language features (Pederson, 2015), which resulted in the highest prediction performance. All approaches found that it was harder to distinguish between PTSD and depression versus detecting the presence of either condition (compared to controls).

On a similar shared dataset, prediction of anxiety was improved (Benton, Mitchell,

& Hovy, 2017) by taking into account gender and 10 comorbid (co-occurring) conditions. Other studies used psychological dictionaries (Linguistic Inquiry and Word Count; LIWC (Pennebaker, Booth, & Francis, 2007) to characterize differences between mental illness conditions (Coopersmith, Dredze, Harman, & Hollingshead, 2015)), or study such difference through building supervised topic models (clusters of semantically-related words) (Resnik et al., 2015).

While a shared dataset has the virtue of allowing for comparison between different approaches, its downside is that sampling and selection biases present in the dataset can affect several studies. On the same dataset, it was observed (Preotiuc-Pietro et al., 2015) that just estimating the age of users a language-based prediction model adequately distinguished between users who had self-declared a PTSD diagnosis and those who had not, and that the language predictive of a self-declared diagnosis of depression and PTSD had a large overlap with the language predictive of personality. This suggests that it may be users with a particular personality or demographic profile who chose to share their mental health diagnosis on Twitter. This concern may limit the generalizability of results obtained on this dataset.

Prediction based on Forum Membership

Internet-based forums, or discussion websites, offer a space in which users can post about their often stigmatized mental health problems openly. Three studies considered specific mental-health forums.

In Bagroy, Kumaraguru, & De Choudhury (2017), forum (reddit) posts were used to study the mental well-being of U.S. university students. A prediction model was trained on data gathered from reddit mental health support communities and applied to the posts collected from 109 university forums (subreddits) to estimate the level of distress at the universities. Longitudinal analysis suggests that the proportion of mental health posts increases over the course of the academic year, particularly for universities with the quarter system. In general, well-being is lower in universities with more females, lower tuition, and in those located in rural or suburban areas**.** In Gkotsis et al. (2016), the language of 16 different forums (subreddits) covering a range of mental health problems was characterized using LIWC and other markers of sentence complexity.

In De Choudhury, Kiciman, Dredze, Coppersmith, & Kumar (2016), posts of a group of reddit users who posted about mental health concerns were studied and then shifted to discuss suicidal ideation in the future. Several features predicted such a shift: heightened self-focus, poor linguistic style matching with the community, reduced social engagement, and expressions of hopelessness, anxiety, impulsiveness, and loneliness. The prediction model could identify these characteristics with an F-score (the harmonic mean of precision and recall) of .80.

Table 1.

Prediction performances achieved by different mental illness studies reviewed in this paper, along with the dataset, features and prediction settings used.

Ref.	Dataset			Mental Illness Criteria	Features						Prediction Setting	Model	Metric	Performance
	Platform	N (users)	Cases (conditions)		n-grams	LIWC	Sentiment	Topics	Metadata	Others				
Preotiuc-Pietro et al., 2015	Twitter	1,957	Depression = 483 PTSD = 370	self-declared	Y	Y	Y	Y		Age, Gender, Personality	Binary	Logistic Regression	AUC	Depression = .85 PTSD = .91
Padrez et al., 2015	Twitter	900	Depression = 326	self-declared	Y					Bag of Words	Binary	Naive Bayes	AUC	0.94
Benton, Mitchell, & Hovey, 2017	Twitter	9,611	4820 (across 8 Conditions)	self-declared	Y					Gender	Multi-Task	Neural Network	AUC	Depression = .76 Bipolar = .75 Depression = .76 Suicide Attempt = .83
Nadeem, 2016	Twitter	21,866	11,866 (across 4 Conditions)	self-declared	Y	Y	Y		Y	Tweet Stats	Binary	Log linear classifier	Precision*	Depression = .48 Bipolar = .64 PTSD = .67 SAD = .42
Coppersmith et al., 2015	Twitter	4,026	2013 (across 10 Conditions)	self-declared	Y	Y					Binary	(not reported)	Precision*	Depression = .48 Bipolar = .63 Anxiety = .85 Eating Dis. = .76
Coppersmith, Dredze, & Harman, 2014	Twitter	5,972	PTSD = 244	self-declared	Y	Y					Binary	(not reported)	ROC	(AUC not reported)
Coppersmith, Harman, & Dredze, 2014	Twitter	250	Suicide Attempt = 125	self-declared	Y		Y		Y	Tweet Stats	Binary	(not reported)	Precision*	.70
Schwartz et al., 2014	Facebook	28,749	(continous Depression score)	survey (Personality)	Y	Y		Y		Season	Continuous	Ridge Regression	Correlation	.38
Tsugawa et al., 2015	Twitter	209	Depression = 81	survey (CESD)	Y	Y	Y	Y	Y	Tweet Stats	Binary	SVM	Accuracy	.69
Reece et al., 2016	Twitter	378	Depression = 105 PTSD = 63	survey (CESD)		Y	Y		Y	Time-Series, LabMT	Binary	Random Forests	AUC	Depression = .87 PTSD = .89
De Choudhury et al., 2014	Facebook	165	Post-partum Depression = 28	survey (PHQ-9)		Y	Y		Y	Stats, Social Capital	Binary	Logistic Regression	pseudo-R2**	.36
De Choudhury et al., 2013	Twitter	476	Depression = 171	survey (CES-D + BDI)		Y	Y		Y	network stats	Binary	PCA, SVM w/ RBF kernel	Accuracy	.72

Note. AUC: Area Under the Receiver Operating Characteristic (ROC) Curve; SVM: Support Vector Machines; PCA: Principal Component Analysis. *Precision with 10% False Alarms; **within-sample (not cross-validated); ***using the Depression facet of the Neuroticism factor measured by the International Personality Item Pool (IPIP) proxy to the NEO-PI-R Personality Inventory (Goldberg, 1999).

Analysis and Prediction based on Annotated Posts

Although most studies are computationally focused, annotation studies that involve manually labeling text, can improve understanding of how mental illness is discussed on social media and can supplement computational approaches (Hwang & Hollingshead, 2016; Kern et al., 2016). Most annotation studies on depression focus on identifying posts in which users are discussing their own experience with depression

(Cavazos-Rehg et al., 2016). Annotators are provided with guidelines for how to recognize a broad range of symptoms of depression (Mowery, Bryan, & Conway, 2015) that are derived from clinical assessment manuals such as the DSM-5 (APA, 2013), or a reduced set, such as *depressed mood*, *disturbed sleep* and *fatigue* (Mowery, Bryan, & Conway, 2015). Annotation has also been used to differentiate between mentions of mental illness for the purpose of stigmatization or insult as opposed to voicing support or sharing useful information with those suffering from a mental illness (Hwang & Hollingshead, 2016).

Ethical Questions

The prediction performances of the studies reviewed above suggest that some mental illnesses can indeed be inferred with some accuracy from public (Twitter and forums) or semi-public (Facebook) social media data. While these efforts have generally been motivated by efforts to detect mental illness for the purpose of delivering mental health services, the success of these algorithms raise several ethical questions.

From the perspective of privacy concerns, employers and insurance companies, for example, may be motivated to derive this information. As mental illnesses carry social stigma, data protection and ownership frameworks are needed to make sure the data is not used against the users' interest (McKee, 2013). Few users realize the amount of mental-health-related information that can be gleaned from their digital traces, so transparency about which indicators are derived by whom for what purpose should be part of ethical and policy discourse.

From a mental health perspective, clear guidelines will be necessary to scaffold decision making regarding when algorithmic identifications of severe distress or the potential for self-harm mandate the alerting of mental health providers. There are also open questions around the impact of mis-classifications, and how derived mental health indicators can be responsibly integrated into systems of care (Inkster, Stillwell, Kosinski, & Jones, 2016). Discussions around issues such as these should include clinicians, computer scientists, lawyers, ethicists, and policy makers.

Recommendations for Future Studies

While the studies reviewed here provide some initial insights regarding the state of the science of detecting mental illness on social media, this remains a young field. Several studies have considered changes in the posting behavior in the context of psychopathology, but future studies should combine both online and offline data in order to follow manifestations of psychopathology in the offline world (Inkster, Stillwell, Kosinski, & Jones, 2016). Additionally, social media data should complement more uninterrupted data streams, such as text messages and emails, or always-on sensor data (Mohr, Zhang, & Schueller, 2017).

It will also be useful to integrate social media data collection within large scale cohort studies. Technological advances have made this prospect increasingly attainable. First studies that combine the collection of social media data with medical records are one promising step in that direction (Padrez et al., 2015).

Conclusion

The studies described here demonstrate that depression and other mental illnesses are detectable on several online environments. Advances in natural language processing are making the prospect of large-scale screening of social media for at-risk individuals a near-future possibility. Ethical and legal questions about data ownership and protection, as well as clinical and operational questions about integration into systems of care should be addressed with urgency.

The previous chapter summarized the recent literature on mental health prediction from social media. The following chapter discusses a particular study that used Facebook to predict depression, the most prevalent mental illness. As concluded in the review, all previously published studies used social media (Twitter and Facebook) to predict self-reported depression status, either derived from the users' score on a depression screening survey, or by using keyword searches on Twitter to identify users who declared a depression diagnosis publically. Across both types of studies, the samples are often highly curated and lack ecological validity. The next study seeks to address this shortcoming and for the first time uses depression status established through clinician judgement (as recorded in medical records) as the criterion to be predicted.

Depression has a relatively low base rate in the population (around 20%) for machine-learning prediction tasks, which makes a hard problem to solve algorithmically: After all, a simple decision rule that would declare all subjects free from depression would be correct in 80% of the cases. This establishes a hard base line to beat. As a result, in many studies the samples are rebalanced artificially, to include about as many depressed and non-depressed users which limits the ecological validity of these studies. The study presented in the following tackles the prediction task assuming real-life base rates, preserving the generalizability of the results to real-life settings.

CHAPTER 3

PREDICTING DEPRESSION THROUGH FACEBOOK

Depressive disorders are prevalent, persistent, and resource intense. Within a given year, an estimated 7-26% of the U.S. population experiences depression (Kessler *et al.* 2003; Demyttenaere *et al.* 2004), of whom only 13-49% receive minimally adequate treatment (Wang et al., 2005). By 2030, unipolar depressive disorders are predicted to be the leading cause of disability in high income countries (Mathers and Loncar, 2006). The U.S. Preventive Services Task Force recommended screening adults for depression in circumstances in which an accurate diagnosis, treatment, and follow-up can be offered (O'Connor *et al.* 2009). These high rates of underdiagnosis and undertreatment suggest that existing procedures for screening and identifying depressed patients are inadequate. There is a need and opportunity for the development of novel methods to screen for patients suffering from depressive disorders.

Using patient's Facebook language data, we built an algorithm to predict the first appearance of a diagnosis of depression in the medical records of a sample of patients presenting to a single, urban emergency department. Previous research has demonstrated the feasibility of using Twitter (De Choudhury, Gamon, Counts, & Horvitz, 2013b; Reece et al., 2016) and Facebook language and activity data to predict depression (Schwartz et al., 2014), postpartum depression (De Choudhury, Counts, Horvitz, & Hoff, 2014), suicidality (e.g., Homan et al., 2014), and post-traumatic stress disorder (e.g., Coppersmith, Harman, & Dredze, 2014b), relying on self-report of diagnoses on Twitter (Coopersmith, Dredze, Harman, Hollingshead, & Mitchell, 2015; Pedersen, 2015) or the participants' responses to screening surveys (De Choudhury et al., 2013b; De Choudhury et al., 2014; Reece et al., 2016) to establish participants' mental health status. This study is the first to use social media data to predict clinical diagnoses not based on self-report but medical records and thus clinician-assessment.

As described in Padrez et al. (2015), patients were approached in an urban academic Emergency Department (ED) and consented to share their own Facebook statuses shared on their profiles ("wall") and access to their medical records. We use mentions of depression-related ICD codes in patients' medical records as a proxy for clinical assessment of depression, which Trinh et al. suggest is feasible with moderate accuracy (2011). 114 patients had a diagnosis of depression in their medical records. For these patients, we determined the date at which the first such diagnosis was recorded in the Electronic Medical Record of the hospital system, and only included Facebook data generated by the user before this date. We sought to realistically model the application of a Facebook-based algorithm applied to patients presenting consecutively in a Primary Care setting by matching every depressed patient with five non-depressed control patients who we simulated presented to the ED on the same day as the depressed user (and had thus generated Facebook data in the same time-span), for a total sample of 683 patients (depression base rate 1:5, or 16.7%).

Materials and Methods

Participant recruitment and data collection. This study was approved by the Institutional Review Board at the University of Pennsylvania. The flow of the data collection is described in Padrez et al. (2015). In total, 11,224 patients were approached in the emergency department over a 26-month period. Patients were excluded if they were under 18 years old, suffered from severe trauma, were incoherent, or demonstrated evidence of severe illness. Of these, 2,903 agreed to share both their social media data and their electronic medical records (EMRs), which resulted in 2,679 (92%) unique EMRs. 1,175 patients (44%) were able to log in to their Facebook accounts and our Facebook app was able to retrieve any Facebook posting language up to 6 years prior, ranging from July 2008 through September 2015.

From the health system's EMRs, we retrieved demographics (age, sex, and race) and prior diagnoses (by International Classification of Diseases [ICD-9] codes). We considered patients as depressed if their EMRs mentioned ICD codes 296.2 (Major Depression) or 311 (Depressive disorder, not elsewhere classified), resulting in 180 patients with any Facebook language (base rate 180 / 1,175 = 15.3%, or 1:5.53). Of the 180 depressed patients, 114 patients (63%) had at least 500 words in status updates preceding their first recorded diagnosis of depression.

To model the application in a medical setting and control for annual patterns in depression, we randomly matched every depressed patient with 5 non-depressed patients who had at least 500 words in status updates preceding the same day as the first recorded diagnosis of depression of the patient they were "control patients" for, yielding a sample of 114 + 5x114 = 684 patients[4]. We excluded one patient from the sample for having less than 500 words after excluding unicode tokens (such as

[4] We excluded 40 users with any Facebook language from the set of possible controls if they did not have the above ICD codes but only depression-like diagnoses that were not temporally limited, i.e. recurrent Depression (296.3) or Dysthymic Disorders (300.4), Bipolar disorders (296.4-296.8), Adjustment disorders or PTSD (309). We additionally excluded 36 patients from the possible control group if they had been prescribed any anti-depressants (SSRIs) without having been given an included depression ICD code.

emojis), for a final sample of N = 683 patients.

Sample Descriptives. Sample descriptives are shown in Table 1. Among all 683 patients, the mean age was 29.9 (SD = 8.57); most were female (76.7%) and Black (70.0%). Depressed patients were more likely to have posted more words on Facebook (Difference between medians = 3,794 words, Wilcoxon W = 27,712, $p = 0.014$), and be female ($\chi 2$ (1, N = 583) = 7.18, $p = 0.007$), matching national trends (Rhodes et al. 2001; Kumar et al. 2004; Boudreaux et al. 2008).

Table 1.
Sample Descriptives

	Depressed	Non-Depressed	Sign. difference?
N	114	569	
Mean age (SD)	30.9 (8.1)	29.7 (8.65)	-
% Female	86.8%	74.7%	p = 0.007
% Black	69.1%	75.4%	-
Mean word count (SD)	19,784 (27,736)	14,802 (21,789)	p = 0.072
Median word Count	10,655	6,861	p = 0.014

Note. Differences in age and mean word count were tested for significance using t-tests, % Female and % Black using $\chi 2$-tests with continuity correction, and median words counts using Wilcoxon rank sum test with continuity correction.

Word and phrase extraction. We determined the relative frequency with which users used words (unigrams) and 2-two phrases (bigrams) using our open source Python-based language analysis infrastructure (see dlatk.wwbp.org).

Topic modelling. We modelled 200 topics from the Facebook statuses of all users using an implementation of Latent Dirichlet Allocation (LDA) provided by the MALLET package (McCallum, 2002). LDA semantically clusters words based on co-occurrence--akin to factor analysis--but appropriate for highly non-normal unigram frequency distributions. LDA yields interpretable units of analysis that implicitly disambiguate word senses. After modelling, we derived every users' use of the 200 topics (200 values per user).

Topic presentation. When visualizing the word clouds in Figure 3, we show the top 15 words per topic with the highest probability in that topic; the size of the words within the topic is the rank of this probability. Color shade aids reusability and carries no meaning.

Temporal feature extraction. We split the time of the day into six bins of four hours in length, and for every user calculated which fraction of statuses was posted in these bins. Similarly, we determined the fraction of posts made on different days of the week.

Meta feature extraction. For every user, we determined how many unigrams were posted per year, the average length of the posts (in unigrams), and the average length of unigrams.

Dictionary extraction. Linguistic Inquiry and Word Count (LIWC 2015, Pennebaker et al., 2015) provides dictionaries (lists of words) widely used in psychological research. We matched the extracted unigram frequencies against these dictionaries to determine the users' relative frequency of use of the 73 LIWC dictionaries.

Prediction models. We used machine learning to train predictive models using the unigrams, bigrams and 200 topics, using 10-fold cross-validation to avoid overfitting (similar to Kosinski, Stillwell, & Gaepel, 2013). In this cross-validation procedure, the data is randomly partitioned into 10 stratified folds, keeping depressed users and her five "control users" within the same fold. A L2-penalized (ridge) logistic regression is trained, and evaluated across the remaining fold; the procedure is repeated 10 times, and an out-of-sample probability of depression is estimated for every patient. Varying the threshold of this probability for depression classification uniquely determines a combination of True and False Positives Rates which form the points of a ROC curve. We summarize overall prediction performance as the area under this ROC curve (AUC), which is suitable for describing prediction accuracies with highly unbalanced classes.

Language associations. To determine if a language feature (topic or LIWC category) was associated with (future) depression status, we determined its AUC with future depression status: as these features are continuously valued and depression status is binary, thresholding on different values of the feature frequency for depression classification determines combinations of True Positive and False Positive values, which trace out the points of the ROC curve and yields an AUC for every language feature. To evaluate if a language feature was associated with depression status over and above age, sex and ethnicity, we use within-sample logistic regression to build a demographic base null model (AUC = .62). Based on this model, we use a nonparametric permutation test with a million iterations to create a null distribution of AUCs, and locate the language feature's AUC to this distribution, yielding a p-value.

Controlling for multiple comparisons. In addition to the customary significance thresholds, we also report if a given language feature meets a $p < 0.05$ significance threshold corrected with the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995) for multiple comparisons.

Results

Prediction of Depression

We evaluated the performance of our prediction model in a cross-validation framework, comparing the probability of depression estimated by our algorithm against the actual future mental health status of the patient. Varying the threshold of this probability for diagnosis uniquely determines a combination of True and False Positives Rates which form the points of a ROC curve; overall prediction performance can be summarized as the area under this curve (AUC).

What mattered most in the prediction was the language content of the Facebook posts. To yield interpretable and fine-grained language units of analysis, we extracted 200 language topics using Latent Dirichlet Allocation (LDA), a method akin to factor analysis but appropriate for word frequencies. We trained a language model based on the relative frequencies with which patients expressed these topics, as well as word and 2-word phrases, obtaining an AUC of 0.67.

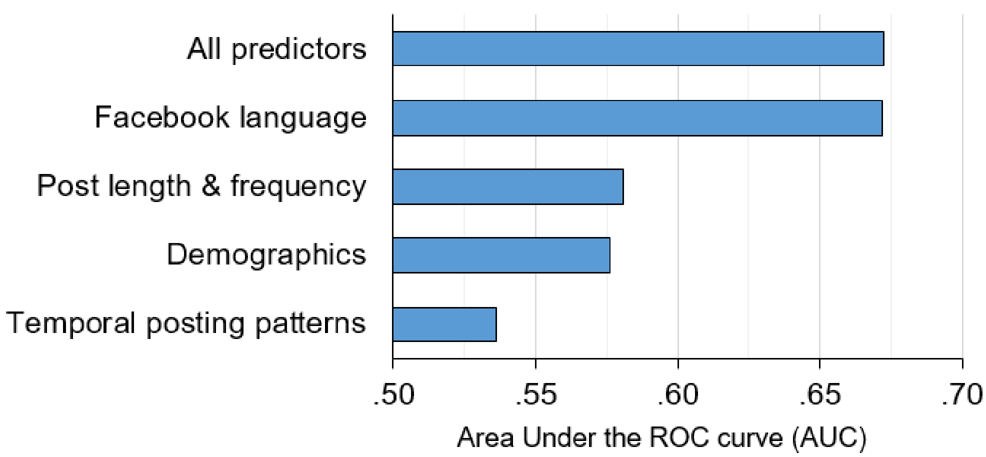


Figure 1. Prediction performances of future depression status based on demographics and Facebook posting activity, reported as cross-validated out-of-sample Areas under the ROC curve (AUCs).

How do these prediction performances compare against other methods of screening for depression? To our knowledge, only one previous study has assessed the concordance of screening surveys with diagnoses of depression recorded in EMRs, as in this study (Noyes[5]) shown in Fig. ROC together with our Facebook model. The results suggest that the Facebook-prediction model obtains screening accuracies comparable to validated self-report depression scales. The relatively stronger performance of our prediction model with laxer thresholds (favoring probability of detection over the probability of false alarms) suggests that

Facebook may best be used as an initial screening method to identify patients for further follow-up either through a self-report survey or clinician assessment.

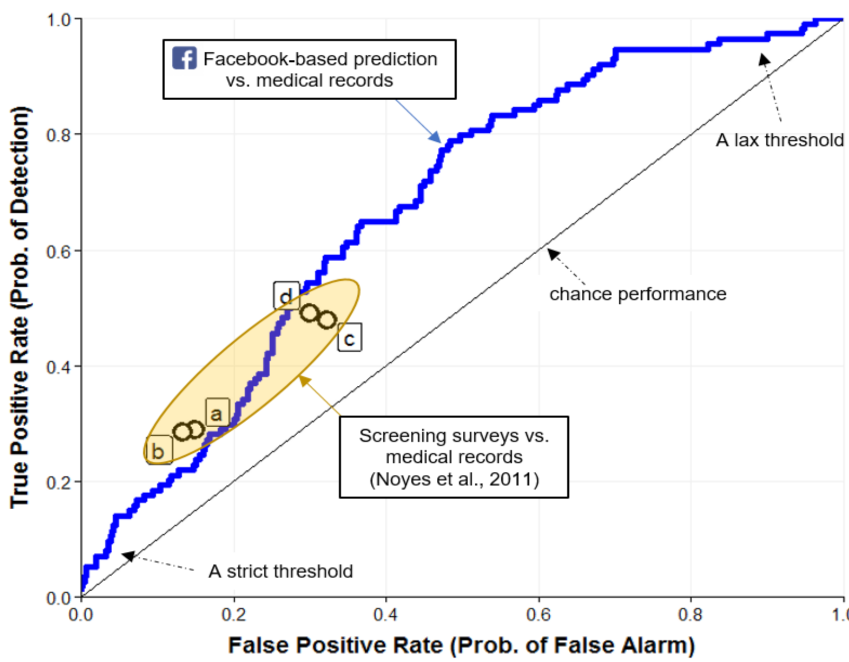


Figure 2. Receiver Operating Characteristic (ROC) curve for a Facebook activity-based prediction model (all predictors combined; blue), and points as combinations of True and False Positive Rates reported by Noyes et al. (2011) for different combinations of

[5] Noyes et al. (2011) sought to benchmark claims data against self-report depression scales as the criterion variable in a sample of N = 1,551 elderly adults; we have derived the points given in Fig. 2 from the confusion matrices they published. They included the ICD-9 used by us (296.2 and 311) among their "extended set" of codes.

depression surveys (**a, b**: Mini-International Neuropsychiatric Interview–Major Depressive Episode Module; **c, d**: Geriatric Depression Scale with a cut-off > 6) and time windows in Medicare claims data (**a, c**: within 6 months before and after survey, **b, d**: within 12 months).

Considering aspects of users' Facebook activity other than language, depressed users only differed modestly from non-depressed users in their temporal post patterns (diurnally and across days of the week; AUC = 0.54), unlike previous work that observed that depressed users are more likely to post during night hours (De Choudhury et al., 2013b). Posting length and frequency (meta-features) contained about as much information about depression status as demographics (both AUC = .58), with the median annual word count across posts being 1,424 words higher for depressed users (Wilcoxon W = 26,594, p = .002). Adding temporal and meta-features to the language-based prediction model did not substantially increase prediction performance, suggesting that the language content captures the depression-related variance in the other feature groups.

Comparison with previous findings. In our sample non-depressed and depressed users were balanced 5:1 to simulate prediction "in the wild." In previous work this balance has been closer to unity (e.g., 1.78:1 in De Choudhury et al., 2013b, 0.94:1 in Reece et al., 2016). When limiting our sample to balanced classes (1:1), we obtain an AUC of 0.68 and F1 score (the harmonic mean of precision and recall) of 0.66, which is comparable to the F1 score of 0.65 reported by Reece et al., (2016) and 0.68 reported by De Choudhury et al. (2013b) based on Twitter data and survey-reported depression. The fact that language content captures the depression-related variance in the other feature groups dovetails with previous work (De Choundhury et al., 2013b, Preotiuc-Pietro et al., 2015).

Language markers of depression. To better understand what language may serve as markers of future depression status, we determined how depressed and non-depressed users differed in their relative frequencies of use of the 200 LDA topics and Linguistic Inquiry and Word Count, an expert crafted dictionary of terms frequently used in psychological research (LIWC 2015, Pennebaker et al., 2015). We controlled for demographics by comparing the within-sample AUCs of models combining these language features with demographic controls against the within-sample AUC = .062 baseline given by a demographic model (age, gender, ethnicity) using a nonparametric permutation test to provide significances.

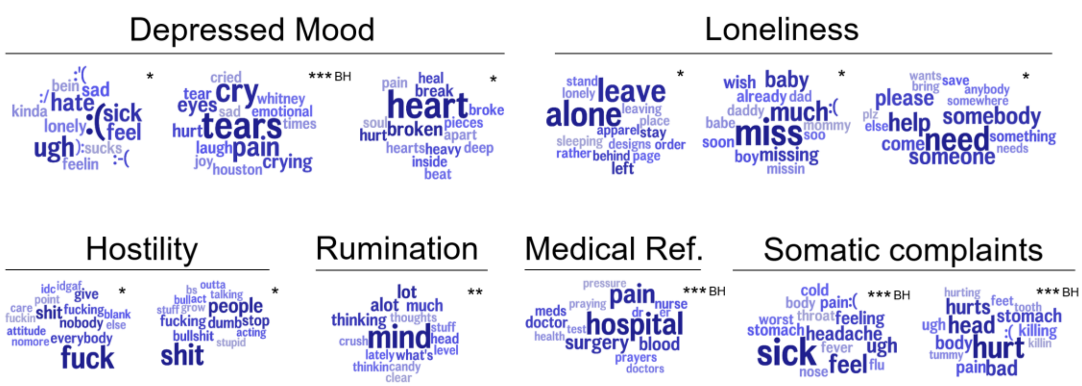


Figure 3. Language topics significantly positively associated by AUC with a future depression diagnosis over and above a baseline AUC of demographic controls. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; ^{BH} $p < 0.05$ after Benjamini-Hochberg correction for multiple comparisons.

We identified 22 (out of 200) topics and 25 (out of 73) LIWC dictionaries as significantly ($p < .05$) positively associated with future depression status over and above the baseline of demographic controls. Figure 3 shows 12 of these topics organized into themes; Table 2 shows the associated LIWC dictionaries.

We observed face-valid emotional language markers of depressed mood (topic: *tears, cry, pain*; AUC = 0.64, $p < 0.001$), loneliness (topic: *alone, leave, left*; AUC = .64, $p = 0.031$) and hostility (topic: *fuck, shit, everybody*; AUC = .64, $p = 0.038$). The LIWC dictionaries negative emotion (AUC = 0.66, p < 0.001; most frequent words: *smh, fuck, hate*) and sadness (AUC = 0.67, p < 0.001; *miss, lost, alone*) captured similar information.

We observed depressed users using more 1st person singular (LIWC dictionary: AUC = .68, p < 0.001; *I, my, me*) and fewer 1st person plural pronouns (LIWC dictionary: AUC = .64, p = 0.014; *we, our, us*), suggesting a preoccupation with the self. 1st person singular pronouns were found by a recent meta-analysis to be one of the most robust language markers of cross-sectional depression status (Edwards & Holtzman, 2017) and by a preliminary longitudinal study of future depression status, as observed in this study (Zimmerman, Brockmeyer, Hunn, Schauenburg, & Wolf, 2016).

Cognitively, depression is thought to be associated with perseveration and rumination, specifically on self-relevant information (Sorg, Vogele, Furka, & Meyer, 2012) which manifests as worry and anxiety when directed towards the future (Edwards & Holtzman, 2017). In line with these conceptualizations, we observed language markers both suggestive of increased rumination (topic: *mind, alot, lot*; AUC = 0.65, $p = 0.002$) and anxiety (LIWC dictionary: AUC = 0.64, $p = 0.013$; *scared, upset, worry*).

Primary care physicians often cite somatic complaints as a frequent feature of depression reported by their patients (Rush, 1993), be it because patients perceive or choose to report somatic symptoms at higher rates (Simon, VonKorff, Piccinelli, Fullerton, & Ormel, 1999). As may be expected given data collection in an Emergency Department, among depressed users we observed language markers of somatic complaints (topic: *hurt, head, bad*; AUC = 0.66, $p < 0.001$; LIWC dictionary: health: AUC = 0.66, $p < 0.001$; *life, tired, sick*). We also observed increased medical references (topic: *hospital, pain, surgery*; AUC = 0.67, $p < 0.001$), depressed individuals are known to be more likely to visit the ED multiple times within a six-month period (Boudreaux *et al.* 2006).

Table 2
LIWC Dictionaries Associated with Depression.

Positively assoc. with dep.	$AUC_{in\text{-}sample}$	Negatively assoc. with dep.	$AUC_{in\text{-}sample}$
Pronouns		**Pronouns**	
1st pers singular (I, me)	.68 ***	1st pers plural (we, our)	.64 *
3rd pers singular (s/he)	.64 *	**Other**	
Emotions		Work	.65 **
Feel (perceptual process)	.68 ***	Hear (perceptual process)	.65 **
Negative Emotions	.66 ***	Power (Drives)	.65 **
Sadness	.67 ***		
Anxiety	.64 *		
Cognitive Processes			
Discrepancy	.66 ***		
Tentative	.65 **		
Other			
Health	.66 ***		
Present focus	.65 **		

Note. Shown here are all pronoun and psychological process LIWC dictionaries significantly associated with future depression status at multiple-comparison corrected significance levels ($p_{BH} < .05$) beyond a baseline of demographic controls (AUC = .62), with strengths of associations given as within-sample Areas under the ROC curve (AUCs). Superordinate dictionaries which include dictionaries shown here (like the Personal Pronoun dictionary) are not shown. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Discussion

Our results show that Facebook-based models do about as well as screening surveys in identifying patients with depression when benchmarked against medical records. The profile of depression-associated language markers is nuanced, covering emotional (sadness, depressed mood), interpersonal (hostility, loneliness) and cognitive processes (self-focus, rumination) which previous research has established as determinants and consequences of depression.

The growth of social media and continuous improvement of machine learning algorithms means that social-media-based screening will become increasingly feasible and more accurate. Being able to identify depressed patients matters, as it touches upon many elements of health care delivery. Depressed patients have increased risk of death from nearly all major medical causes (Zivin, et al., 2015); after diagnosed heart failure, for example, their mortality is increased twofold (Fan et al., 2014). Depressed patients are also more likely to visit the ED multiple times within a six-month period (Boudreaux et al., 2006). Identifying these individuals on their first ED visit would help them connect with necessary care while simultaneously relieving an ED's often scarce resources (American Hospital Association, 2005) of the burden of multiple visits.

Because of its low base rate and varying presentation, depression is hard to detect by primary care physicians: the number of both detected and missed cases can be less than the number of false positives (Inkster, Stillwell, Kosinski, & Jones, 2016). In addition, ED physicians in particular are trained to identify and treat acute over chronic conditions; depression may not be noticed in an emergency setting. This is confirmed by studies that suggest that ED physicians show low sensitivity (< 40%) in their unaided assessment of patient depressive status (Perruche et al. 2011).

Thus, previous research has recommended improving detection through a multi-step assessment processes (Inkster et al., 2016) – our results suggest that Facebook maybe a valuable first step in such a screening procedure. Akin to triaging, a standard ED procedure used to determine severity of symptoms, unobtrusive social media language analysis may offer a preliminary but immediate view of mental health that can be follow up on with existing (more resource-intensive) self-report screening instruments that have demonstrated acceptable sensitivity and specificity when benchmarked against gold-standard clinician-delivered structured clinical interviews (Gilbody, Sheldon, & House, 2008). The combination of Facebook screening and validated screening instruments may yield higher prediction performance than unaided assessment by clinicians.

A single Facebook authorization allows the retroactive collection of data covering multiple years, allowing the clinician to observe the severity of depression over time, and enabling ongoing measurement, affording a longitudinal perspective that self-report measures omit. The language findings across different nuanced symptom clusters suggest that analysis of Facebook may eventually yield a dashboard highlighting specific symptoms to the clinician. Further, prediction models may be calibrated to use different thresholds depending on the use case. With a lax threshold favoring a higher probability of detection, Facebook-based screening may be used to triage patients for further assessment. With a strict threshold favoring a low probability of false alarms, in principle Facebook-based models can be used to screen large populations, and identify the most severe cases for targeted follow up.

With the potential for improved mental health care delivery, these technologies also raise questions about privacy, data protection and data ownership. Few users will realize that they might be disclosing their mental health status to third parties through as simple an act as adding an app on Facebook, which may include insurances or employers.

Clear guidelines are needed on how consumers are to be informed about what information is derived from their data. Developers and policymakers need to address the challenge that the application of an algorithm may change social media posts into protected health information.

While data linking mental health diagnoses with social media is unprecedented, by modern standards of big data research our final sample was relatively small. Still, it already provides empirical evidence that the text-based analysis of social media language can serve as a cost-efficient and efficacious front-line of mental health assessment in real life medical settings. Together with the growing sophistication, scalability and efficacy of technology-supported treatments for depression (Foroushani, Schneider, & Assareh, 2011; Newman, Szkodny, Llera, & Przeworski, 2011), this suggests that both detection and treatment for mental illness may soon meet individuals in the digital spaces they already inhabit.

The preceding three chapters introduced computational linguistic methods and their application to characterize and predict depression, the most prevalent mental illness. In the next chapter, similar methods are employed to characterize and predict atherosclerotic heart disease, the leading cause of death. Across the previous chapters, the objects of the analysis were individuals, and the predominant source of text was Facebook statuses. In the next chapter, using language collected through Twitter, the computational linguistic methods are generalized to the community-level, specifically, to U.S. counties. Starting with a sample of one billion Tweets, the locations of origin were determined and mapped onto U.S. counties. The rest of the analysis is comparable to the preceding chapters: Rather than a person, a U.S. county is now the unit of analysis, and mortality rates from atherosclerotic heart disease are the health outcome being predicted. The successful application of these methods across U.S. counties in the following chapter suggest that social-media-based prediction methods generalize beyond individuals to communities, suggesting that they can offer contributions to epidemiology and public health.

CHAPTER 4

PREDICTING HEART DISEASE THROUGH TWITTER

Heart disease is the leading cause of death worldwide (World Health Organization, 2011). Identifying and addressing key risk factors such as smoking, hypertension, obesity, and physical inactivity has significantly reduced risk (Ford & Capewell, 2011). Psychological characteristics such as depression (Lett et al., 2004) and chronic stress (Menezes, Lavie, Milani, O'Keefe, & Lavie, 2011) have similarly been shown to increase risk through physiological effects (such as chronic sympathetic arousal) and deleterious health behaviors (such as drinking and smoking). On the other hand, positive characteristics such as optimism (Boehm & Kubzansky, 2012) and social support (Tay, Tan, Diener, & Gonzalez, 2013) seem to decrease risk, most likely through similar pathways.

In the 2020 Strategic Impact Goal Statement, the American Heart Association suggests that to further reduce the risk for heart disease, "population-level strategies are essential to shift the entire distribution of risk" (Lloyd-Jones et al., 2010, p. 589). Like individuals, communities have characteristics that contribute to health and disease, such as norms, social connectedness, perceived safety, and environmental stress (Cohen, Farley, & Mason, 2003). One challenge to addressing community-level psychological characteristics is the difficulty of assessment; traditional approaches that use phone surveys and household visits are costly and have limited spatial and temporal precision (Auchincloss, Gebreab, Mair, & Diez Roux, 2012; Chaix, Merlo, Evans, Leal, & Havard, 2009).

Rich information about the psychological states and behaviors of communities is now available in big social media data, offering a flexible and significantly cheaper alternative for assessing community-level psychological characteristics. Social media-based digital epidemiology can support faster response and deeper understanding of public health threats. For example, Google used search queries to measure trends in influenza, providing earlier indication of disease spread than the Centers for Disease Control and Prevention (CDC; Ginsberg et al., 2009). Other studies have used Twitter to track Lyme disease, H1N1, depression, and other common ailments (Chew & Eysenback, 2010; De Choudhury, Counts, & Horvitz, 2013; Paul & Dredze, 2011a; 2011b; Quincy & Kostkova, 2009; Salathé, Freifeld, Mekaru, Tomasulo, & Brownstein, 2013; Seifter, Schwarzwalder, Geis, & Aucott, 2010; St Louis & Zorlu, 2012*)*.

Methods for inferring psychological states through language analysis have a rich history (Pennebaker, Mehl, & Niederhoffer, 2003; Stone, Dunphy, Smith, Ogilvie, 1966). Traditional approaches use “dictionaries” —predetermined lists of words—associated with different constructs (e.g., *sad, glum, crying* are part of a negative emotion dictionary; Pennebaker, Chung, Ireland, Gonzales, & Booth, 2007). *Open-vocabulary* approaches identify predictive words statistically and are not based on traditional dictionaries (Schwartz et al., 2013), offering a complementary approach to language analysis.

In this study, we analyzed social media language to identify community-level psychological characteristics associated with atherosclerotic heart disease (AHD) mortality. In a dataset of tens of millions of Twitter messages (tweets), we used dictionary-based and open-vocabulary analyses to characterize the psychological language correlates of AHD mortality. We also gauged the amount of heart disease-relevant information in Twitter language by building and evaluating predictive models of AHD mortality and compared the language models to alternative models with traditional demographic and socioeconomic risk factors.

Methods

We collected tweets from across the United States, determined their counties of origin, and derived language variables for each county (e.g., the relative frequencies that people from the county expressed anger or engagement). We correlated these county-level language variables with county-level age-adjusted AHD mortality rates obtained from the CDC. To gauge the amount of heart disease-relevant information contained in the Twitter language, we compared the performance of prediction models based on Twitter language against models that contained county level measures of (a) socioeconomic status (income and education), (b) demographics (percentage of Blacks, Hispanics, married, and female residents), and (c) health variables (incidence of diabetes, obesity, smoking, and hypertension). All procedures were approved by the University of Pennsylvania Institutional Review Board.

Data Sources

We used data from 1,347 U.S. counties that had AHD mortality rates, county-level socioeconomic and demographic variables, and at least 50,000 tweeted words. Over 88% of the U.S. population lives in the included counties (U.S. Census Bureau, 2010).[6]

[6] Excluded counties for which heart disease, demographic, and socioeconomic information was available had smaller populations (median population 12,932 in *n* = 1,796 excluded counties vs. 78,265 in included

Twitter data. Twitter messages (tweets) are 140-character messages containing information about emotions, thoughts, behaviors, and other personally salient information. In 2009 and 2010, Twitter made a 10% random sample of tweets (``the Garden Hose'') available for researchers through direct access to their servers. We obtained a sample of 826 million tweets collected between June 2009 and March 2010. Many Twitter users self-reported their locations in their user profiles, which we used to map the tweets to counties (for details, see Automatic County Mapping section in the Supplemental Material available online). This resulted in 148 million county-mapped tweets across 1,347 counties for which a sufficient number of tweets and reliable mortality and demographic data were available.

Heart disease data. Counties are the smallest socioecological level for which most CDC health variables and U.S. Census information are available. From the CDC (2010) we obtained county-level age-adjusted mortality rates for AHD (International Classification of Disease 10 [ICD] code I25.1), which is the single ICD 10 code with the highest overall mortality in the U.S. (prevalence: 52.5 deaths per 100,000). We averaged AHD mortality rates across 2009 and 2010 to match the time period of the Twitter language dataset.

Demographic and health risk factors. From the American Community Survey (2009), we obtained county level high school and college graduation rates, from which we created an index of educational attainment; we also obtained median income and

---

counties), higher rates of AHD (Hedges' g = .48 [.38, .57], n = 597), lower income (g = -.42 [-.53, -.32], n = 496) and education (g = -.61 [-.72, -.51], n = 496). Median age was not significantly different (g = 0.003 [-.08, 0.8], n = 1,004).

percent married. From the U.S. Census Bureau (2010), we obtained percentage of female, Black, and Hispanic residents. From the CDC's Behavioral Risk Factor Surveillance System (2009-2010), we obtained self-reported prevalence of diabetes, obesity, smoking, and hypertension (common cardiovascular risk factors), for which county-level estimates had previously been derived (see Table S1 in Appendix B for detailed source information).

Analytic Procedure

Language variables from Twitter. An automatic process was used to extract the relative frequency of *words* and *phrases* (one to three word sequences) for every county. For example, the relative frequency of the word "hate" ranged from .003% to .240% across counties (see Tokenization in the Supplemental Material available online).

We then derived two more types of language use variables from counties based on the relative word frequencies: (a) predetermined *dictionaries* of psychologically-related words, yielding the relative frequency of words used by counties for the given dictionaries (e.g., *positive emotion* words accounted for 0.5% of all words in a county on average); and (b) 2,000 automatically created *topics* (clusters of semantically-related words; see "Topic Extraction" in the Supplemental Material available online), yielding the probability that each county mentioned each topic. We used pre-established dictionaries for anger, anxiety, positive/negative emotions, positive/negative social relationships, and engagement/disengagement (Pennebaker et al., 2007; Schwartz et al., 2013). Topics were previously automatically derived (Schwartz et al., 2013).

Because words can have multiple senses or can be used in the context of irony or negation, it is important to empirically gauge how well such lists of words measure what is intended (Grimmer & Stewart, 2013). To that end, human raters evaluated the dictionaries to determine that they accurately measured the psychological concept intended. For each of the eight dictionaries, two independent raters examined 200 tweets containing dictionary words and rated whether the word expressed the associated dictionary concept within the tweet. A third rater was brought in to break ties. Judges rated the dictionaries to have accuracies between 55% and 89% (see Table S2 in Appendix B).[7]

Statistical analysis. Dictionary and topic language variables were correlated with county AHD mortality rates using ordinary least squares linear regression. Each language variable was entered individually into the regression equation, and then simultaneously entered with education and income as controls. As 2,000 topics were tested, to avoid type I errors, we applied the Bonferroni-correction to the significance threshold (i.e., for the correlation of one of 2,000 topics to be significant, its *p*-value would have to meet a threshold of $p < .05/2000$, or .000025).

Predictive models. A predictive model of county AHD mortality rates was created based on all of the Twitter language variables – a single model that used the county word, phrase, dictionary, and topic usages as independent variables, and outputted the AHD mortality rate as the dependent variable. We used regularized linear regression ("ridge regression") to fit the model (see "Predictive Models" in the Supplemental Material

[7] The anxiety and positive relationship dictionaries were rated as having the lowest accuracies (55.0% and 55.5% respectively; **see** Table S2), whereas the accuracy of the other dictionaries was markedly higher (average accuracy 82.1%). Cross-correlations of dictionaries (Table S3 in Appendix B) revealed that the positive relationship and the anxiety dictionaries unexpectedly were positively correlated with all other dictionaries.

available online). Similarly, we created predictive models of county AHD mortality rates based on different combinations of Twitter language, county demographic (percentage of Blacks, Hispanics, married, and female residents), socioeconomic (income, education), and health variables (incidence of diabetes, obesity, smoking, and hypertension).

We avoided distorted results (due to model "overfitting" —picking up patterns simply due to chance) by using a 10-fold cross-validation process which compared model predictions to out-of-sample data. The predictive models were created by fitting the independent variables to the dependent variable (AHD mortality) on a random 9/10th of the counties (the training set), and then evaluated on the remaining 1/10th (hold-out set). We evaluated the models by comparing the actual CDC-reported mortality rates with each models' predicted rates using a Pearson product-moment correlation. The procedure was repeated ten times, once for each tenth of the counties, and then averaged together for an overall prediction performance across all counties. To compare predictive performance between two models, we conducted paired *t*-tests comparing the sizes of standardized residuals of county-level predictions from each model.

Results

Dictionaries. Anger, negative relationships, negative emotions, and disengagement significantly correlated with greater age-adjusted AHD mortality (Pearson r = .10 [95% confidence interval = .05, .16]. to .17 [.11, .22]; Table 1). After controlling for SES (income and education), all five negative factors (including anxiety) were significant risk factors for AHD mortality ($r_{partial}$ = .06 [.00, .11] to .12 [.07, .17]), suggesting that Twitter language captures information not accounted for by SES. Positive emotions and engagement were associated with lower AHD mortality (r = -.11 [-.17, -

.06] and -.16 [-.21, -.10] respectively). Engagement remained significantly protective after controlling for SES (r_{partial} = -.09 [-.14, -.04]); positive emotion was marginally significant ($r_{partial}$ = -.05 [-.00, -.11]). The positive relationships dictionary[8] showed a nonsignificant association with AHD mortality (r = .02 [-.04, .07]).

[8] The word "love" was removed from the dictionary, as it accounted for more than a third of all word occurrences in the dictionary, and distorted the results (see discussion).

Table 1
Correlations Across 1,347 Counties Between Atherosclerotic Heart Disease (AHD) Mortality and Twitter Language Measured by Dictionaries.

	Twitter Language as Measured by Dictionaries	**Correlation with Atherosclerotic Heart Disease Mortality** (Pearson r with 95% confidence intervals)
Risk Factors	Anger	.17 [.11, .22] ***
	Negative Relationships	.16 [.11, .21] ***
	Negative Emotions	.10 [.05, .16] ***
	Disengagement	.14 [.08, .19] ***
	Anxiety	.05 [.00, .11] †
Protective Factors	Positive Relationships[3]	.02 [-.04, .07]
	Positive Emotions	-.11 [-.17; -.06] ***
	Engagement	-.16 [-.21, -.10] ***

Note. Anger and anxiety come from LIWC dictionaries (Pennebaker et al., 2007); others are our own (Schwartz et al., 2013). Positive correlations indicate higher AHD mortality.
*** $p < 0.001$; † $p < 0.10$.

Topics. We complemented the dictionaries with an open-vocabulary approach, using automatically created topics that form semantically-coherent groups of words, calculating each county's probability of mentioning each topic, and correlating topic use with AHD. Figure 1 shows 18 topics that were significantly correlated with AHD mortality.[9] For risk factors, we observed themes of hostility and aggression (*sh*t, *sshole, f***ing*; r = .18 [.12, .23] to .27 [.22, .32]), hate and interpersonal tension (*jealous, drama, hate*; r = .16 [.11, .21] to .21 [.16, .26]), and boredom and fatigue

[9] We grouped topics into seemingly related sets, and added labels to summarize our sense of the topics. These labels are open to interpretation, and we present the most prevalent words within the topics for inspection. County-level topic and dictionary frequency data can be downloaded from wwbp.org.

(*bored, tired, bed*; r = .18 [.12, .23] to .20 [.15, .25]). After controlling for SES, seven of the nine risk topics remained significant at Bonferroni-corrected levels ($r_{partial}$ = .12 [.07, .17] to .25 [.20, .30], $p < 7 \times 10^{-6}$).

For protective factors, topics about positive experiences (*wonderful, great, hope*; r = -.14 [-.19, -.08] to -.15 [-.21, -.10]) related to lower mortality, mirroring the dictionary-based results. A number of topics reflected skilled occupations (*service, skills, conference*; r = -.14 [-.20, -.09] to -.17 [-.22, -.12]). One set of topics reflected optimism (*hope, opportunities, overcome; r* = -.12 [-.18, -.07] to -.13 [-.18, -.07]), which has demonstrated robust associations with reduced cardiovascular disease risk at the individual level (Boehm & Kubzansky, 2012; Chida & Steptoe, 2008). After controlling for SES, the protective topics (Figure 1 bottom) were significant at the traditional $p < .05$ level, but were no longer significant at Bonferroni-corrected levels.

Prediction. Figure 2 compares the predictions of AHD mortality from regression models with several independent variables. Combining Twitter and ~~the~~ ten traditional demographic, SES and health predictors slightly but significantly increased predictive performance over a model that only included the ten traditional predictors ($r_{twitter\_demo\_SES\_health}$ = .42 [.38, .46], $\mathrm{r}_{demo\_SES\_health}$ = .36 [.29, .43]; t(1,346) = -2.22; p = .026), suggesting that Twitter has incremental predictive validity over and above traditional risk factors. A predictive model using *only* Twitter language performed slightly better than a model using the ten traditional factors ($r_{twitter}$= .42 [.38, .45], t(1,346) = -1.97, p = .049).

To explore these associations in greater detail, Table S4 (Appendix B) compares the performance of prediction models containing stepwise combinations of Twitter and sets of demographic (percentage of Blacks, Hispanics, married and female residents), socioeconomic (income and education), and health predictors (incidence of diabetes, obesity, smoking and hypertension). For all combinations of sets of traditional predictors, adding Twitter significantly improves predictive performance ($t(1346) > 3.00$, $p < 0.001$). Adding traditional sets of predictors to Twitter in no case significantly improved predictive performance.

Taken together, these results suggest that the AHD-relevant variance in the ten predictors overlaps with the AHD-relevant variance in the Twitter language features, suggesting that Twitter may be a marker for these variables, while also having incremental predictive validity. Figure 3 shows CDC-reported 2009-2010 AHD mortality (left) and Twitter predicted mortality (right) for the densely populated counties in the Northeastern U.S.; a high degree of overlap is evident.

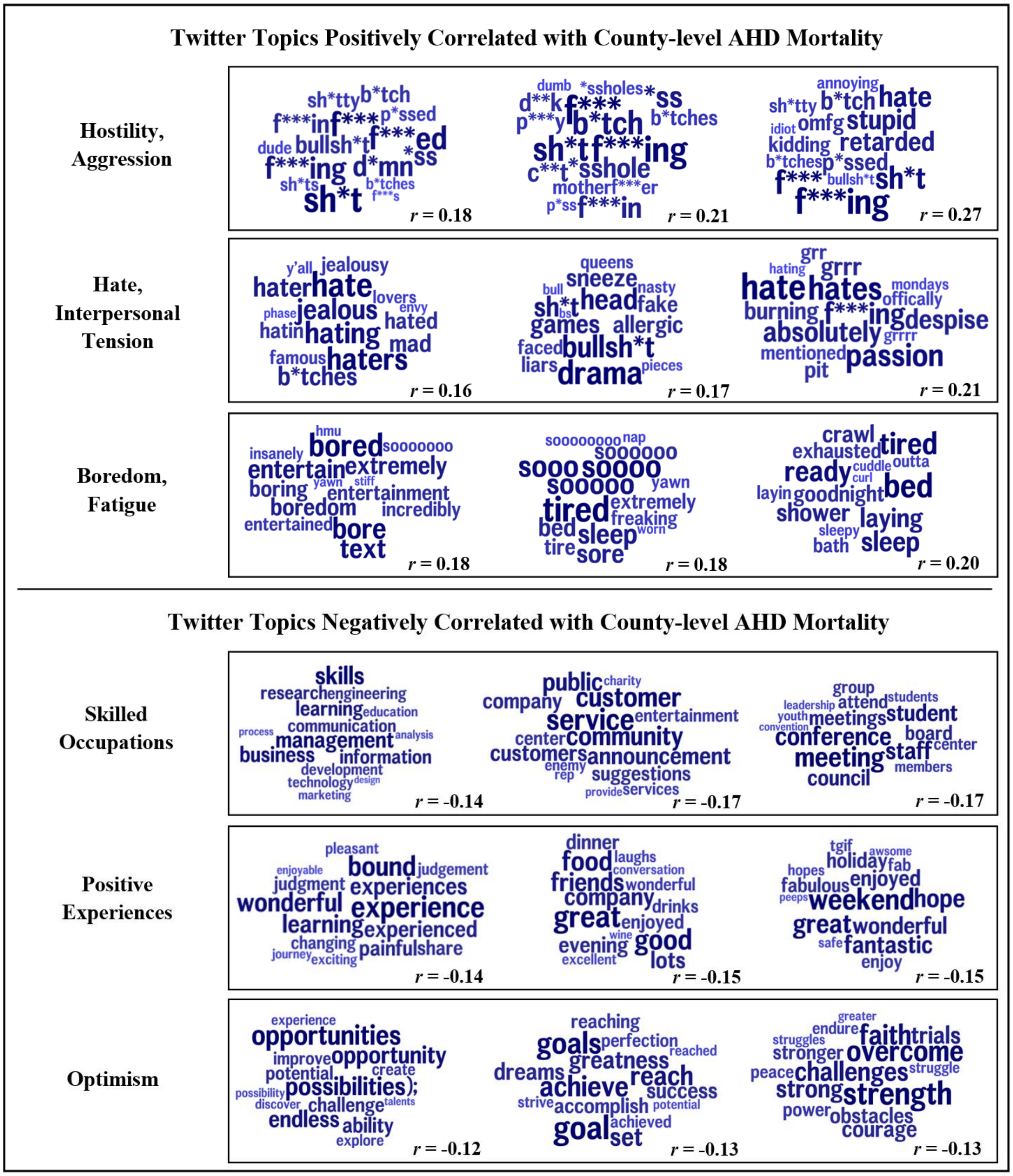


Figure 1. Twitter topics most correlated with age-adjusted AHD mortality (significant at a Bonferroni-corrected significance level of $p < 2.5 \times 10^{-5}$). The size of the word represents its prevalence within the topic (larger = more prevalent; see Supplemental Material available online for details).

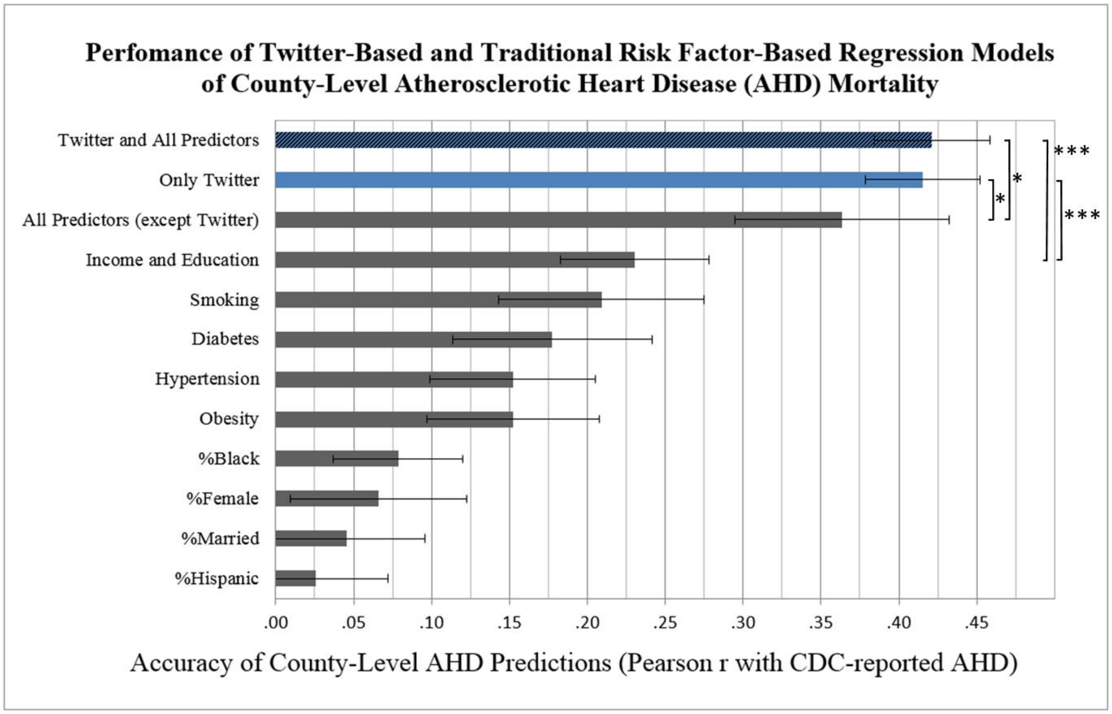


Figure 2. Performance of regression models predicting age-adjusted atherosclerotic heart disease (AHD) mortality from Twitter language, compared to SES, health, and demographic variables, and a combined model (higher values mean better predictions; error bars show 95% confidence intervals). The model is trained on one part of the data ("training set") and evaluated on another ("hold-out set"), to avoid distorted accuracies due to chance ("overfitting"). A model combining Twitter and all predictors significantly outpredicted the model with all predictors (combining all SES, demographic, and health variables), suggesting that Twitter has incremental predictive validity. Twitter language by itself significantly out-predicted a model with all SES, demographic, and health predictors. *** $p < 0.001$; * $p < 0.05$.

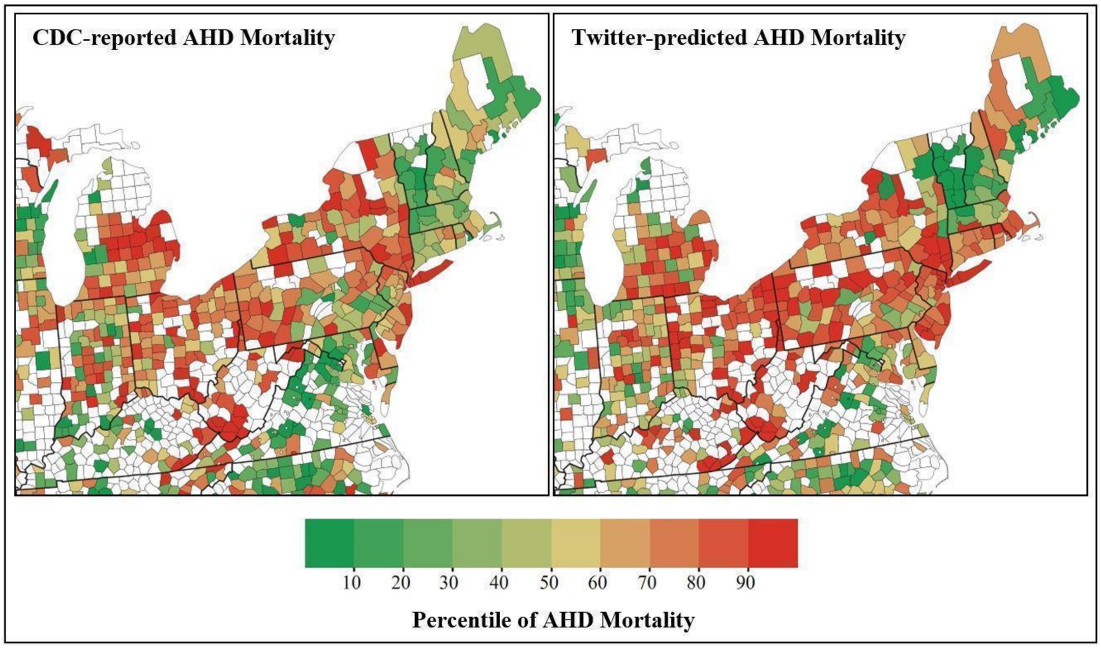


Figure 3. Map of Northeastern U.S. counties showing age-adjusted rates of atherosclerotic heart disease (AHD) mortality as reported by the CDC (left), and estimated through the Twitter-language-only prediction model (right). The counties were randomly split into a "training" and a "hold out set." The Twitter model is trained on the training set and predictions are made on the hold out set, to avoid distorted accuracies due to chance ("overfitting"). This procedure is repeated to derive predictions for all counties, shown here. Red counties have higher rates of mortality, green lower. White counties indicate that reliable CDC or Twitter language data were unavailable.

Discussion

Our study had three major findings. First, language expressed on Twitter revealed several community-level psychological characteristics that were significantly associated with atherosclerotic heart disease (AHD) mortality risk. Second, positive emotions and engagement were protective from AHD mortality risk, whereas negative emotions (especially anger), disengagement, and negative relationships were risky. Third, our predictive results suggest that the information contained in Twitter fully accounts for—and adds to—the AHD-relevant information in ten representatively-assessed demographic, socioeconomic, and health variables.

Taken together, our results suggest that language on Twitter can provide plausible indicators of community-level psychosocial health that may complement other spatial methods used in epidemiology (c.f. Auchincloss et al., 2012), and that these indicators are associated with risk for cardiovascular mortality.

Our findings point to a community psychological risk profile similar to risk profiles that have been observed at the individual level. County-level associations between AHD mortality and negative emotions (relative risk[10] [RR] = 1.22), anger (RR = 1.41), and anxiety (RR = 1.11) were comparable to individual level meta-analytic effect sizes for depressed mood (RR = 1.49; Rugulies, 2002), anger (RR = 1.22; Chida & Steptoe, 2009), and anxiety (RR = 1.48; Roest, Martens, de Jonge, & Denollet, 2010).

While less is known about the protective effects of positive psychological variables at the individual level, our findings align with a growing body of research supporting the cardiovascular health benefits of psychological well-being (Boehm & Kubzansky, in press). Engagement, which has long been considered an important component of successful aging (Rowe & Kahn, 1987), emerged as the strongest protective factor. Positive emotions were also protective, in line with numerous reviews that find positive emotions to be protective from illness and disease (e.g., Howell, Kern, & Lyubomirsky, 2007; Pressman & Cohen, 2005). Fredrickson and colleagues (2000) have argued that positive emotions may undo the negative cardiovascular aftereffects of anxiety-induced cardiovascular reactivity. Optimism has demonstrated relatively robust association with reduced risk of cardiovascular events at the individual level (Boehm & Kubzansky, 2012; Chida & Steptoe, 2008). Demonstrating the value of data-driven

---

[10] To compare our findings with published effect sizes, correlation coefficients were converted to relative risk following Rosenthal and DiMatteo (2001).

language analyses, we did not have a predefined optimism dictionary, but our topic analyses seemingly identified this protective factor, as indicated by topics containing *hope, opportunities, overcome* (Figure 1, bottom).

Overall, our topic findings were similar to and converged with our theory-based dictionary results (cross-correlations are given in Supplemental Table S3 in Appendix B). While theory-based findings can be more easily tied to existing literature, topic analyses provide a richer portrait of specific behaviors and attitudes (e.g., cursing, frustration, being tired) that correspond to broad psychological characteristics (such as anger or stress) associated with an increased risk for AHD mortality. Data-driven analyses like topics may help identify novel psychological, social, and behavioral correlates of disease.

With theory-based dictionaries, results can be driven by a few frequent but ambiguous words. For example, the original positive relationships dictionary (Schwartz et al., 2013) was surprisingly associated with increased risk, as was its most frequent word, *love. Love* accounted for more than a third of the total usage of the positive relationships dictionary (5.3 million occurrences of *love* compared to 15.0 million for the entire dictionary), effectively driving the dictionary results. Reading through a random sample of tweets containing “love” revealed them to be mostly statements about loving *things*, not people[11]. Excluding *love* from the dictionary reduced the correlation between the positive relationship dictionary and heart disease from r = .08 [.03, .13] to a non-

[11] In addition to this word sense ambiguity, a factor analysis of the words in the positive relationships dictionary revealed two factors with opposing correlations to socioeconomic status (SES; income and education). A general social factor (*friends, agree, loved*) correlated with higher SES (r = .14), and a ‘partnership' factor (*relationship, boyfriend, girlfriend*) with lower SES (r = -.43) and higher AHD mortality (r = .18). Love loaded much higher on this second factor (see Table S5 in Appendix B). Love may be picking up on the fact that in lower SES areas users share more about personal relationships, thus distorting the original positive relationship results.

significant *r* = .02 [-.04, .07].

These results demonstrate the pitfalls of interpreting dictionary-based results at face value, and underscore the importance of interpreting dictionary-based results in light of the most frequent words contained in the dictionaries which can drive the overall dictionary results in unexpected ways. For transparency, we have included the correlations with AHD for the 10 most frequent words across the eight dictionaries in Table S6 in the Supplemental Material available online. These findings also highlight the value of triangulating language analyses across different levels of analysis (words, topics, dictionaries) for more robust interpretations.

Given that the typical Twitter user is younger (median age is 31; Fox, Zickurh, & Smith, 2009) than those at risk for AHD, it is not obvious why Twitter should track heart disease mortality. The people tweeting are not the people dying. However, the tweets of younger adults may disclose characteristics of their community, reflecting the shared economic, physical, and psychological environment. At the individual level, multiple pathways connect psychological variables and heart disease risk, including health behaviors, social relationships, situation selection, and physiological reactivity (Friedman & Kern, 2014). These pathways occur within a broader social context, which directly and indirectly influence the individual's life experiences. Local communities create physical and social environments that influence the behaviors, stress experiences, and health of its members (Diez Roux & Mair, 2010; Lochner, Kawachi, Brennan, & Buka, 2003). Epidemiological studies have found that the aggregated characteristics of communities, such as social cohesion and social capital, account for a significant portion of variation in health outcomes, independent of individual characteristics (Leyland, 2005; Riva, Gauvin,

& Barnett, 2007), such that the combined psychological character of the community is more informative for predicting risk than are the reports of any one individual. The language of Twitter may be a window into the aggregated and powerful effects of the community context.

Our study has several limitations. Twitter messages constitute a biased sample in two ways. First, Twitter messages may reflect social desirability biases as people manage their online identity (Rost, Barkhuus, Cramer, & Brown, 2013). Second, Twitter users are not representative of the general population. The Twitter population tends to be more urban and have higher education (Mislove, Lehmann, Ahn, Onnela, & Rosenquist, 2011). In 2009, the Twitter median age of 31 (Fox et al., 2009) was 5.8 years below the U.S. median age (U.S. Census Bureau, 2010). Our Twitter-based prediction model outperforms models based on classical risk factors in predicting AHD mortality; this suggests that, in spite of the biases, Twitter captures as much unbiased AHD-relevant information about the general population as traditional, representatively-assessed predictors.

Third, our findings are cross-sectional; future research should address the stability of psychological characteristics of counties across time. Fourth, we relied on AHD mortality rates reported as underlying causes of death on death certificates by the CDC, based on coding practices which may be inconsistent (Pierce & Denison, 2010). Finally, language associations do not point to causality; language on social media may complement other epidemiological methods, but causal inferences from observational studies have been repeatedly noted (Diez Roux & Mair, 2010).

Traditional approaches for collecting psychosocial variables of large representative samples, such as the CDC's Behavioral Risk Factor Surveillance System and Gallup polls, tend to be expensive, based on merely thousands of people, and are often limited to a minimal, predefined list of psychological constructs. A Twitter-based system to track psychosocial variables is relatively inexpensive, and can potentially generate estimates based on tens of millions of people with much higher resolution in time and space. It is comparatively easy to create dictionaries automatically for different psychological or social constructs, allowing the testing of novel hypotheses. Our approach opens the door to a new generation of psychological informational epidemiology (Eysenbach, 2009; Labarthe, 2010), and could bring us closer to understanding what community-level psychological factors are important for the cardiovascular health of communities and should become the focus of intervention.

GENERAL DISCUSSION

In the first chapter, three dictionary-based ("closed-vocabulary") programs for text analysis (the General Inquirer, DICTION, and Linguistic Inquiry and Word Count) were compared with two "open-vocabulary" methods (topic modelling through Latent Dirichlet Allocation [LDA] and Differential Language Analysis) across 13 million status updates from 65,000 Facebook users. While the psychological insights gained through closed and open-vocabulary methods were similar, data-driven open-vocabulary results were more specific and useful for psychological hypothesis generation. In addition, the comparative performance from cross-validated machine learning prediction models suggests that encoding users' language as distributions over 2,000 LDA topics captured more variance related to demographics and personality than dictionaries.

The second chapter reviews studies (mostly published in computer science) that use the methods introduced in the first chapter to predict mental illness from social media language. These studies suggest that depression and other mental illnesses are detectable in several online environments, particularly on Facebook, Twitter and in web forums. While this suggests that the analysis of social media text may allow for the screening of mental illness, the ecological validity of existing studies is limited. Firstly, most studies use depression status determined through screening surveys or public sharing of a diagnosis on Twitter as the criterion, as opposed to clinician judgement. Secondly, the existing studies rarely include an appropriate balance of depressed to non-depressed users in their samples which would resemble the low depression base rate observed in real-life settings.

The third chapter presents a study designed to demonstrate the feasibility of using Facebook data to screen for depression, alleviating some of these methodological concerns. Facebook data was collected in conjunction with access to electronic medical records in the Emergency Department of a large urban teaching hospital. To simulate screening through Facebook, only Facebook data preceding the first recorded diagnosis of depression in the medical record was used in prediction models, with a depression base rate of 17% in the sample. Facebook-based prediction models were able to predict future depression with fair accuracy, and did about as well as screening surveys in identifying patients with depression when benchmarked against medical records in another study. The language associated with depression dovetails with existing conceptualizations of depression covering emotional (sadness, depressed mood), interpersonal (hostility, loneliness) and cognitive processes (self-focus, rumination). This study is the first demonstration of language analysis of social media as a screening tool for depression in a real-world medical setting.

In the fourth chapter, the application of social media text analysis is generalized to the community level and applied to characterize and predict mortality from atherosclerotic heart disease, the leading cause of death. Rather than Facebook data as in the preceding chapters, public Twitter data is "geo-tagged" to their U.S. counties of origin, yielding county-level language samples. An analysis of the language profiles associated with heart disease using both closed and open-vocabulary approaches reveals negative emotions (especially hostility), disengagement and negative relationships to be associated with increased risk, while positive emotions and engagement showed protective associations. A Twitter-language-based prediction model outperformed a model including ten demographic, socioeconomic and health risk factors (including smoking, obesity, hypertension and diabetes rates), suggesting that Twitter captures variance in heart disease mortality not captured by the traditional variables.

Taken together, the results presented suggest that large scale analysis of social media using methods of natural language processing are a feasible and desirable technology to improve the measurement of population health. In mental health, the findings suggest that Facebook and Twitter can be used to screen for depression in medical settings and identify individuals for further follow-up. A generalization of these methods to measure community-level depression rates seems highly plausible, as suggested by first studies (e.g., De Choudhury et al., 2013a). For physical health, these methods have demonstrated predictive validity in estimating the atherosclerotic heart disease mortality rates across U.S. counties, roughly matching the prediction performance of gold standard epidemiological models.

An analysis of associated social media language yields profiles of psychological risk factors for both depression and heart disease that capture many of the known psychological predictors. Depression appears associated with not just depressed mood but loneliness, hostility and rumination, while heart disease is associated with hostility, negative emotions and disengagement as risk factors, and positive emotions and engagement as protective factors. In this way large scale analysis of social media text can add a "dashboard" of associated psychological processes to our understanding of population health challenges, making no theoretical assumptions *a priori*. This suggests that these methods have the power to identify psychological determinants of population health factors that other approaches may have missed, while simultaneously being able to measure their relative importance. Accordingly, this work has clear applications for both public health and public policy.

For public health—in addition to the contributions discussed above—these technologies suggest that "primordial risk" is now measurable—the psychological risk factors (like stress, or hostility) that lead to negative health behaviors (like overeating, or excessive drinking) that then in turn affect physical health outcomes. In addition, these technologies allow for the data-driven discovery and measurement of positive psychological health assets (like positive relationships, or optimism)—about which relatively less is known—that buffer against negative health outcomes.

For public policy, these technologies suggest that psychological states of large populations can be measured directly, with little temporal lag and high spatial resolution. This method of psychological measurement brings us one step closer to observing the desired outcomes of policy interventions. When, for example, the changes in stress levels of a community in response to changes to walkways and urban greening can be reliably and immediately determined, it will be much easier to make the case that these interventions work, without having to wait for years to observe trends in obesity rates. In this way, large scale analysis of social media can "close to loop" for policy makers, not only by helping to identify determinants of population health, but also by providing a real-time measurement infrastructure to track the psychological impact of policy interventions.

Limitations & Future Directions

There are some ways that deserve attention in which analyses of social media text through methods of computational linguistics have not yet fully matured.

Causality

Very few of the published studies that use analysis of social media to predict outcomes of interest are in the position to make causal claims about the nature of the associated language findings. Most of the studies are cross-sectional; a few have embraced minimal longitudinal designs in which the predictors precede the occurrence of a condition of interest (as in the last chapter of this dissertation, or as in De Choudhury et al., 2013b). Submitting psychological predictors of health outcomes to tests of Granger causality, for example, seems like an obvious direction for future study designs; as do data collection efforts that accompany experimental study designs.

Aggregate vs. Individual-level Prediction

The methods discussed in this dissertation to carry out psychological measurement through social media appear to be strongest when applied in aggregate, for example, at the county-level. This may in part simply be because the aggregation smooths and stabilizes the notoriously sparse distributions of language features, in addition to reducing the reporting and measurement error in the outcome measure (like mortality rates). However, it may also be due to the fact that some associations are stronger at the community than at the individual level—for example, Lawless and Lucas (2011) suggest that the aggregate education level of a community is a stronger predictor of one's life satisfaction than one's own education level, suggesting that the education level of a community encodes more than merely college completion rates.

At the individual level, while Park et al. (2014) and Youyou, Kosinski, & Stillwell (2015) have shown that social-media based predictions can match or exceed the predictions of observer-report when compared with self-report inventories, the accuracy of out-of-sample predictions rarely exceed accuracies of $r = 0.3$ to 0.4 with the outcome of interest. While psychologists are used to observing correlations of this magnitude between psychological traits and measurable behaviors (language use can be thought of as a behavior), the fact that such models account for less than 20% of the variance (R^2) in the outcome ought to caution us about our use of these prediction models to make assessments about individuals in high stakes situations (say for insurance coverage, or loan decisions). In some scenarios, this noisiness of the predictions can be alleviated through proper use and calibration of these technologies in a larger assessment context, for example, by using social media predictions with lax thresholds as a first step in a multi-step screening procedure. However, current capabilities warrant caution about individual-level assessments.

Social Media Biases

Perhaps the most consistent question-objection raised when presenting this research over the years is the question about the biases inherent in using social media data. The major points of concern are *sampling* and *desirability biases*. Sampling biases refer to the concern that social media samples are not fully representative of the population. Self-presentation or desirability biases capture the idea that social media users are sharing updates about the self in part to garner a desired response from their social media audience, be it admiration or social support, and that what they share is in part shaped or limited by the response they hope their content will elicit. Both concerns are justified; I will offer a general response to both concerns before addressing them in turn.

In general, out-of-sample prediction accuracies built over representative outcomes offer an empirical way to establish an upper bound of how much these (and other) biases may distort our findings. The fact that Twitter-language-based prediction models outperform gold standard epidemiological models in predicting population (not narrow sample) mortality rates establishes that--whatever the biases may be that affect the signal captured in Twitter and contribute to noise--they still leave enough signal in the Twitter data to capture a part of the variance large enough for us to take very seriously (e.g., Eichstaedt et al., 2015). Given that the prediction models and outcome data (mortality rates recorded through death certificates) cover more than 80% of the U.S. population, it appears that the predictions of these models generalize to whole populations.

Sampling biases. When machine learning prediction models calibrate themselves in the process of predicting representative data, they will appropriately weigh features to approximate the representative data; in other words, even when using data from a biased sample, they are re-stratifying their coefficients appropriately in the process.

However, not using representative outcome data but only outcome data from users who reach a sufficient threshold of words to be included in a language sample (and thus oversampling users who are frequent posters) may somewhat distort the composition of the sample. When we compared personality traits and demographics in a large Facebook sample (N = 68,264) against users with insufficient Facebook language for analysis, we observed users included for language analysis to skew slightly more introverted (by about a fifth of a standard deviation) and female (66%) (Park et al., 2016). These are small effects, and generally taken care of through statistical control in the exploratory language analyses.

In our experience, the extent of biased sampling in social media is often overestimated by naïve audiences. The median age of Twitter users, for example, only differs by 4 years from the median age of the US population and African-American users are *over*sampled on Twitter (Fox, Zickuhr, & Smith, 2009). And finally, whatever these sampling biases may be, they are rapidly diminishing as social media are used by more and more of the US and global population – in the same way in which limiting samples to smartphones users once raised concerns about introducing sampling biases, while today 77% of the population carry smartphones (Mobile Fact Sheet, 2017).

Social-desirability biases. In addition to the general response offered above regarding the demonstrated accuracy when predicting representative outcomes, even in samples into which social-media users self-select, we have seen no meaningful evidence of social desirability biases distorting our analyses across numerous investigations (Kern et al., 2014a; Kern et al., 2014b; Park et al., 2015; Schwartz et al., 2013b). We were able to predict less desirable traits (like Neuroticism) about as well as desirable ones (like Agreeableness; Park et al., 2015). We have seen highly undesirable psychological characteristics (like hostility or mental illness) emerge as some of the strongest correlates of personality traits. Very often, however, the frequency of occurrence of these undesirable language markers is low, suggesting that undesirable disclosures are less frequent, but that their pattern of covariance with outcomes like personality is preserved. In other words, social media samples may have to be larger to detect highly undesirable traits (to the order of N = 10,000), but their detection is not in principle precluded by the nature of social media.

Ethical Implications

The predictive power of computational linguistic analyses, combined with their relative novelty, raises several ethical concerns. Large percentages of the world's population are now plugged into social media and regularly sharing a large amount of personal information. While people may know that what they share is publicly or semi-publicly accessible, they often do not realize what can be predicted through non-obvious aspects of their writing. For example, algorithms can predict one's gender, political affiliation, sexual orientation, ethnicity, personality, and many other traits with non-negligible accuracy – without the individual ever explicitly mentioning any of these traits (Youyou et al., 2014; Park et al., 2015).

In many ways, the fact that these technologies allow for the micro-targeting of advertisements creates the economic base for the social media ecosystems to exist—advertisement is the business model, and most users seem to tacitly accept this reality.

More concerning are cases in which insurance agencies and financial service firms use this information to assess risk at the individual level. Besides the inherent noisiness of using these methods to generate individual-level predictions, in such high stakes circumstances civil liberties enter into the equation. One could imagine being denied health coverage for their children due to one's Facebook posts, or having one's car premiums raised after a Facebook-based prediction algorithm has inferred one's risk seeking personality trait. A recent attempt by a car insurance provider to use social media data to inform policy pricing in the UK caused a public uproar (Ruddick, 2016), but such publicity cannot always be counted on.

Perhaps the most concerning case involves totalitarian regimes using these methods to control populations. When political affiliation can be inferred, members of opposition political parties could be identified and targeted. Other forms of cultural oppression could also be enacted through these means, such as a repressive regime identifying likely homosexual individuals, for example.

Therefore, given these significant ethical issues, I propose that entities involved in analyzing our “digital footprints” ought to be required to disclose which data they hold, and how they are using it. Google Dashboard, for example, provides such a functionality for Google users (see http://www.google.com/settings/dashboard). Regulators ought to coordinate the legal response to these challenges, and citizens’ rights to their data and transparency about how their data is being used ought to become a *digital* human right in the 21st century. Transnational legislative bodies where appropriate (like the European Union) are likely the most suitable source of internationally coordinated, harmonized and enforceable legislation.

However, ethical issues also arise from failing to take these technologies seriously and failing to make appropriate use of them. When even the strictest thresholds on a prediction algorithm suggests that a Twitter user is severely depressed, questions arise how systems of care can and ought to respond appropriately, and at which point reporting ought to be mandated, and to whom. Perhaps the biggest challenge with using these technologies to identify physically and mentally ill individuals is not the detection itself, but how to design systems of care that can respond appropriately and at scale.

Conclusion

In the early days of a new technology, its use tends to be largely *skeuomorphic*: It recreates the results and aims of old technologies, in ways that are better in some ways. In its simplest form, big data psychology looks similar to psychology as usual, but with overwhelming statistical power because of its many observations—but few researchers can get excited about very small standard errors. However, by adding methods from natural language processing and thereby unlocking the high-dimensional variable space of language, this statistical power has allowed us to siphon the language signal from the noise and create simple and intuitive summaries of the emotional, cognitive and behavioral correlates of any given construct. Soon, a single question related to a proposed new construct answered by a thousand Twitter users may quickly yield the behavioral, emotional and cognitive aspects of the proposed construct, and in one fell swoop shine a searchlight over its nomological net and bootstrap a year's worth of focus groups and participant interviews. Using prediction algorithms built off self-report surveys on a few thousand participants, we can approximate the assessment of millions of people by applying the prediction model to larger language samples, as if they had all taken noisy self-report surveys.

These advancements are certainly laudable—but in my view do not yet represent the potential in the fully matured application of these technologies. The methodological leap of big data psychology requires corresponding conceptual advances and technological integration for us to see the true value of this revolution. For example, one day soon computational linguistic analysis may yield tailor-made cognitive feedback in CBT and prediction algorithms will fine-tune psychological interventions in ways that feel natural and surprisingly thoughtful. The next generation of big data psychology will require technical finesse, but even more so, imagination.

APPENDIX A

Table S1 - *Top Five Word Frequencies within LIWC2015 dictionaries across 12.7 million Facebook statuses*

Dictionary Name	Word 1		Word 2		Word 3		Word 4		Word 5		Word Count
Total function words	the	6.1%	i	6.0%	to	5.8%	a	4.5%	and	4.4%	102,796,062
Total pronouns	i	20.1%	you	10.6%	my	10.0%	it	7.2%	me	5.9%	30,827,244
Personal pronouns	i	27.9%	you	14.7%	my	13.9%	me	8.2%	your	4.6%	22,203,369
1st pers singular	i	48.0%	my	23.8%	me	14.0%	i'm	6.1%	im	3.3%	12,925,539
1st pers plural	we	52.4%	our	18.0%	us	13.5%	lets	4.1%	let's	4.0%	1,369,851
2nd person	you	60.7%	your	19.1%	u	9.0%	you're	3.0%	ur	2.1%	5,364,596
3rd pers singular	her	24.7%	he	22.5%	she	17.0%	his	15.6%	him	11.1%	1,752,055
3rd pers plural	they	62.1%	them	34.3%	themselves	1.8%	they'll	0.8%	they've	0.6%	791,328
Impersonal pronouns	it	25.6%	that	18.1%	this	14.9%	what	9.3%	who	6.2%	8,603,061
Articles	the	55.3%	a	40.8%	an	3.9%					11,364,639
Prepositions	to	24.5%	of	10.5%	in	9.9%	for	9.0%	on	6.4%	24,485,825
Auxiliary verbs	is	17.3%	be	7.5%	have	6.5%	are	4.8%	was	4.4%	18,431,193
Common Adverbs	so	12.6%	just	10.4%	now	6.3%	when	5.5%	back	4.3%	11,181,373
Conjunctions	and	39.5%	so	12.5%	but	9.4%	if	7.4%	when	5.4%	11,344,620
Negations	not	23.8%	no	14.3%	don't	11.3%	can't	8.0%	never	7.5%	4,069,632
Common verbs	is	8.6%	be	3.7%	have	3.2%	are	2.4%	was	2.2%	37,116,279
Common adjectives	as	5.2%	more	4.3%	happy	4.0%	new	3.9%	great	2.9%	9,973,426
Comparisons	like	22.4%	as	13.5%	more	11.0%	best	6.3%	better	6.3%	3,860,228
Interrogatives	what	24.9%	when	18.9%	who	16.5%	how	14.4%	why	9.4%	3,217,748
Numbers	one	48.2%	first	16.1%	two	10.9%	once	5.5%	half	4.7%	1,337,454
Quantifiers	all	28.0%	some	11.5%	more	10.5%	much	8.6%	any	4.1%	4,028,386
Affective processes	:)	7.0%	love	6.7%	good	4.9%	lol	3.1%	happy	3.0%	13,233,081
Positive emotion	:)	9.8%	love	9.3%	good	6.8%	lol	4.3%	happy	4.2%	9,520,167
Negative emotion	:(	8.2%	hate	5.1%	miss	4.7%	bad	4.7%	sick	3.4%	3,608,901
Anxiety	fear	11.5%	scared	9.6%	afraid	9.1%	worry	8.7%	confused	7.5%	250,095
Anger	hate	19.5%	fuck	11.1%	hell	9.6%	stupid	9.1%	sucks	5.0%	936,735
Sadness	miss	18.7%	lost	11.2%	sad	7.9%	sorry	6.9%	alone	6.6%	908,728
Social processes	you	20.3%	your	6.4%	love	5.5%	we	4.5%	who	3.3%	16,021,961
Family	family	18.6%	baby	18.1%	mom	12.3%	dad	7.0%	son	4.6%	873,242
Friends	friends	45.1%	friend	22.2%	dear	9.0%	date	5.0%	honey	2.0%	613,506
Female references	her	32.2%	she	22.1%	girl	8.7%	mom	8.0%	she's	3.2%	1,343,034
Male references	he	24.0%	his	16.6%	man	12.5%	him	11.8%	boy	4.0%	1,643,485
Cognitive processes	all	5.8%	but	5.5%	not	5.0%	if	4.3%	know	3.0%	19,442,918
Insight	know	16.9%	think	10.8%	feel	8.1%	find	4.3%	feeling	3.8%	3,401,616
Causation	how	20.3%	make	14.9%	why	13.3%	because	9.4%	made	7.6%	2,279,300
Discrepancy	if	23.4%	want	11.1%	need	10.1%	would	8.2%	should	5.8%	3,604,900
Tentative	if	19.5%	or	12.0%	some	10.7%	hope	4.4%	any	3.8%	4,326,477
Certainty	all	39.8%	never	10.7%	ever	8.2%	always	7.1%	every	5.3%	2,834,521
Differentiation	but	18.1%	not	16.5%	if	14.4%	or	8.8%	really	6.7%	5,868,107
Perceptual processes	see	9.7%	feel	6.0%	say	5.9%	watching	3.3%	look	3.2%	4,582,865
See	see	20.8%	watching	7.0%	look	6.8%	looking	5.9%	watch	5.2%	2,129,875
Hear	say	29.8%	said	13.3%	says	8.0%	hear	6.6%	listening	4.8%	909,624
Feel	feel	20.3%	feeling	9.5%	hard	8.8%	cold	6.1%	hot	6.0%	1,353,456

Biological processes	love	19.3%	life	11.0%	sleep	4.7%	tired	3.5%	heart	3.5%	4,552,822
Body	sleep	15.6%	heart	11.4%	head	8.2%	face	7.2%	ass	5.1%	1,382,361
Health	life	36.9%	tired	11.7%	sick	9.1%	live	8.8%	pain	4.8%	1,351,578
Sexual	fuck	46.9%	gay	11.0%	sex	8.9%	sexy	8.5%	dick	3.7%	221,229
Ingestion	eat	9.8%	sweet	9.0%	water	6.2%	eating	6.0%	drunk	3.6%	684,483
Drives	love	6.7%	up	6.4%	get	5.9%	we	5.4%	good	4.9%	13,243,757
Affiliation	love	22.2%	we	18.1%	friends	7.0%	our	6.2%	us	4.7%	3,960,514
Achievement	work	18.7%	best	10.5%	better	10.4%	first	9.2%	lost	4.4%	2,336,086
Power	up	23.7%	god	7.6%	over	7.3%	down	7.0%	best	6.9%	3,542,321
Reward	get	19.4%	good	16.2%	got	14.5%	great	7.3%	best	6.1%	4,003,996
Risk	bad	26.0%	stop	18.3%	wrong	11.7%	worst	5.6%	lose	5.4%	653,043
Past focus	was	12.9%	got	9.2%	had	7.8%	been	4.8%	done	3.0%	6,349,857
Present focus	is	12.0%	be	5.2%	have	4.5%	are	3.4%	do	3.0%	26,566,687
Future focus	will	18.1%	going	14.2%	then	11.6%	gonna	5.8%	hope	5.0%	3,854,342
Relativity	in	9.5%	on	6.1%	at	4.1%	up	3.3%	out	3.1%	25,628,205
Motion	go	15.0%	going	11.6%	come	5.7%	gonna	4.8%	put	3.3%	4,732,118
Space	in	19.8%	on	12.6%	at	8.5%	up	6.8%	out	6.4%	12,316,596
Time	now	7.7%	when	6.6%	back	5.2%	then	4.8%	night	4.5%	9,198,466
Work	work	25.7%	school	13.2%	class	6.0%	working	5.1%	read	3.0%	1,698,965
Leisure	fun	16.2%	family	11.5%	facebook	10.5%	play	7.1%	playing	4.5%	1,412,751
Home	home	34.6%	family	16.1%	bed	15.9%	house	15.6%	room	7.2%	1,006,049
Money	free	14.8%	worth	9.8%	spend	7.0%	spent	6.4%	bought	5.1%	540,764
Religion	god	44.7%	hell	15.0%	pray	6.7%	soul	5.9%	holy	3.9%	600,742
Death	die	26.5%	dead	21.0%	died	12.5%	alive	11.6%	war	9.7%	226,593
Informal language	:)	13.8%	u	7.2%	lol	6.1%	well	4.7%	:(	4.4%	6,755,972
Swear words	fuck	16.0%	hell	13.9%	ass	10.9%	sucks	7.3%	crap	6.4%	648,603
Netspeak	:)	20.4%	u	10.6%	lol	9.1%	:(	6.5%	gonna	5.0%	4,544,205
Assent	awesome	21.5%	ok	13.7%	yeah	13.3%	yes	11.8%	cool	8.6%	717,289
Nonfluencies	well	43.0%	oh	32.1%	ugh	10.7%	sigh	4.5%	ah	3.8%	729,465
Fillers	blah	47.0%	idk	30.4%	dunno	9.0%	whoa	7.9%	woah	5.2%	68,630

APPENDIX B

Table S1 *Variable Sources and Transformation*

Variable			Description of variable	Unit	Years covered	Source
Included variable	**Transformation**	**Categories**				
Atherosclerotic Heart Disease (AHD) mortality	averaged across years		International Classification of Disease (ICD) 10 code I25.1 recorded as underlying cause of death on death certificates, prepared for the county level and age-adjusted through the CDC (using year 2000 population estimates)	per 100,000 population	2009-2010	CDC Wonder, Underlying Cause of Death (CDC, 2010)
Income	log-transformed		Median household income	2010 inflation-adjusted US dollars	2008-2010	American Community Survey (ACS, 2010) 3-Year Estimates (Table DP03)
Educational Attainment Index	Independently standardized and then averaged	High school grad	Attainment of high school graduation or higher	% of population	2008-2010	ACS 3-Year Estimates (Table DP02) (ACS, 2010)
		College grad	Attainment of bachelor's degree or higher			
Diabetes			Adults (age 20+) diagnosed with diabetes	% of population	2008-2010	County-level estimates based on CDC's Behavioral Risk Factor Surveillance System (BRFSS) data (2009-2010), obtained through 2013 County Health Rankings (CHR; 2010) (see note).
Obesity			Body Mass Index >= 30, based on self-reported height and weight			
Smoking			Current adult smokers who have smoked >= 100 cigarettes in their lifetime		2005-2011	
Hypertension	averaged	male	Male adults (age 30+) who self-reported systolic BP of at least 140mm Hg and/or self-reported taking medication	% of population	2009	County-level estimates prepared through the Institute for Health Metrics and Evaluation (IHME; 2009) on the basis of CDC BRFSS data (see note).
		female	Female adults (age 30+) who self-reported systolic BP of at least 140mm Hg and/or self-reported taking medication			
% Black			Population of one race - Black or African American alone	% of population	2010	U.S. Census, Demographic Profile Data (Table DP01) (U.S. Census Bureau, 2010)
% Hispanic			Hispanic or Latino			
% Female			Female			
% Married	averaged	male	Male adults (age 15+) now married (not separated)	% of population	2008-2010	ACS 3-Year Estimates (Table DP02) (ACS, 2010)
		female	Female adults (age 15+) now married (not separated)			

Note on sources used for selected variables:

Diabetes and Obesity: County Health Rankings (CHR; 2010) used data from the National Center for Chronic Disease Prevention and Health Promotion's Division of Diabetes Translation (part of the CDC), which provides the Diabetes Public Health Resource (DPHR; 2010). DPHR used data from the CDC's Behavioral Risk Factor Surveillance System (BRFSS; 2009-2010), an ongoing national survey. DPHR developed county-level estimates from state-level BRFSS data using small area estimation techniques, including Bayesian multilevel modeling, multilevel logistic regression models, and a Markov Chain Monte Carlo simulation method.

Smoking: County-level estimates (based on BRFSS state-level data) were calculated for CHR by CDC staff.

Hypertension: The Institute for Health Metrics and Evaluation (IHME; 2009) used National Health Examination and Nutrition Survey data (1999-2008) to characterize the relationship between self-reported and physical measurements for various health factors. They used the resulting model to predict physical measurements for 2009 BRFSS participants (who supplied self-reported measures) and employed small area estimation techniques to estimate hypertension prevalence at the county-level.

Table S2

Dictionary Evaluation

	Dictionary	Top Ten Dictionary Words by Frequency	Two Rater Agreement	Accuracy
Risk Factors	Anger	shit f*** hate damn b*tch hell f***ing mad stupid b*tches	70.0%	60.0%
	Negative Relationships	hate alone jealous blame evil rude lonely independent hated ban	86.0%	75.5%
	Negative Emotion	sorry mad sad scared p*ssed crying horrible afraid terrible upset	87.0%	79.5%
	Disengagement	tired bored sleepy lazy blah meh exhausted yawn distracted boredom	91.0%	88.0%
	Anxiety	crazy pressure worry scared awkward scary fear doubt horrible afraid	81.5%	55.0%
Protective Factors	Positive Relationships	love home friends friend team social welcome together kind dear	75.0%	55.5%
	Positive Emotion	great happy cool awesome amazing glad excited super enjoy wonderful	93.0%	88.5%
	Engagement	learn interesting awake interested alive learning creative alert involved careful	74.5%	79.0%

Note. Each dictionary was evaluated by two independent raters. 200 random instances of tweets containing words from the dictionary in question were extracted, and the expert raters determined whether the word expressed the associated dictionary concept within the tweet. On average, the raters agreed 81.5% of the time, and a third rater was brought in to break ties. Accuracy refers to the percentage of tweets that expressed the associated dictionary concept, out of the 200 random instances sampled for every dictionary.

Table S3

Cross-Correlations between Dictionaries and Topics

		Anger	Negative Relation-ships	Negative Emotion	Disengage-ment	Anxiety	Positive Relation-ships†	Positive Emotion	Engage-ment
Anger		1	.76 [.73, .78]	.60 [.57, .64]	.72 [.69, .74]	.29 [.24, .34]	.18 [.26, .36]	-.33 [-.38, -.28]	-.30 [-.35, -.25]
Negative Relationships				.70 [.68, .73]	.67 [.64, .70]	.37 [.32, .41]	.42 [.50, .58]	-.04 [-.09, .01]	-.09 [-.14, -.04]
Negative Emotion					.55 [.51, .59]	.43 [.38, .47]	.45 [.50, .58]	.19 [.14, .24]	.04 [-.02, .09]
Disengagement						.29 [.24, .34]	.28 [.37, .46]	-.16 [-.21, -.11]	-.27 [-.32, -.22]
Anxiety							.38 [.29, .39]	.23 [.18, .28]	.16 [.11, .21]
Positive Relationships								.48 [.43, .52]	.23 [.18, .28]
Positive Emotion									.61 [.58, .64]
Topics	**Included Word**								
Hostility, Aggression	bullsh*t	.94	.58	.43	.62	.19	-.03	-.45	-.40
	a**hole	.93	.62	.48	.61	.19	.00	-.41	-.39
	retarded	.81	.65	.56	.54	.21	.06	-.26	-.30
Hate, Inter-personal Tensions	hating	.88	.74	.54	.68	.23	.13	-.33	-.36
	drama	.87	.67	.53	.66	.26	.18	-.28	-.29
	passion	.67	.84	.66	.60	.33	.37	.02	-.08
Boredom, Fatigue	bored	.70	.60	.47	.87	.20	.16	-.26	-.35
	tired	.69	.70	.62	.87	.31	.32	-.04	-.21
	bed	.50	.61	.56	.69	.30	.41	.08	-.12
Skilled Occupations	management	-.42	-.32	-.23	-.41	.03	.29	.38	.69
	service	-.41	-.28	-.17	-.39	.08	.33	.51	.63
	conference	-.45	-.28	-.16	-.42	.11	.34	.56	.65
Positive Experiences	experience	-.30	-.12	-.01	-.26	.15	.42	.57	.76
	company	-.30	-.12	.11	-.21	.18	.54	.78	.67
	weekend	-.35	-.11	.09	-.22	.14	.55	.89	.62
Optimism, Resilience	opportunities	-.33	-.20	-.12	-.31	.10	.35	.41	.69
	achieve	-.21	-.07	.00	-.22	.17	.36	.39	.68
	strength	-.14	.06	.04	-.08	.29	.55	.48	.68

Note. Dictionary cross-correlations (Pearson *r*) are given, with 95% confidence intervals in brackets. To ease inspection, topic-dictionary correlations are color formatted, ranging from dark red (strongly negative) to dark green (strongly positive). Particularly strong correlations between topic clusters and dictionaries are emphasized with bolder boxes. Topics correspond to the topics shown in Figure 1, in the same order. The "included words" are dominant unique words in each cloud, which help identify the topic.
† The word "love" was removed from the dictionary, as it accounted for more than a third of all word occurrences in the dictionary, and distorted the results (see discussion).

Table S4

Performance of Regression Models Predicting AHD Mortality on the Basis of Different Sets of Predictors

Model	Demographic	SES	Health	Twitter	Accuracy of County-Level AHD Prediction	
1	X				.14 [.09, .19]	*** (1 vs 2)
2	X			X	.42 [.38, .45]	
3		X			.23 [.18, .28]	*** (3 vs 4)
4		X		X	.41 [.38, .45]	
5			X		.27 [.20, .34]	*** (5 vs 6)
6			X	X	.42 [.38, .46]	
7	X	X			.32 [.27, .37]	*** (7 vs 8)
8	X	X		X	.41 [.38, .45]	
9	X		X		.33 [.26, .40]	*** (9 vs 10)
10	X		X	X	.42 [.38, .46]	
11		X	X		.29 [.23, .35]	*** (11 vs 12)
12		X	X	X	.42 [.38, .46]	
13	X	X	X		.36 [.29, .43]	* (13 vs 14); * (13 vs 15)
14	X	X	X	X	.42 [.38, .46]	
15				X	.42 [.38, .45]	

Note. Performance of regression models predicting atherosclerotic heart disease (AHD) mortality from demographic variables (percentage of Blacks, Hispanics, married, and female residents), socioeconomic variables (income and education), health variables (incidence of diabetes, obesity, smoking, and hypertension), Twitter language, and all combinations of these sets of predictors. Accuracy refers to the Pearson *r* correlation between the set of predictors and CDC reported AHD. Brackets give 95% confidence intervals. The models are trained on one part of the data ("training set") and evaluated on another ("hold-out set"), to avoid distortion through chance. A model combining Twitter and all predictors (Model #14) significantly outpredicted the model with all predictors (Model 13), suggesting that Twitter has incremental predictive validity. Twitter language by itself significantly outpredicted a model with all SES, demographic and health predictors (Model 15 compared to Model 13). Predictive performance between two models was compared through paired t-tests, comparing the sizes of standardized residuals of county-level predictions from each model. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; † $p < 0.10$.

Table S5

Varimax-rotated Factor Structure of the County-level Frequencies of the 20 most Frequent Words in the Positive Relationship Dictionary

Words	Partnership factor	Social factor
love	**.65**	.39
home	.11	.35
friends	.47	**.53**
friend	.43	**.48**
team	-.07	.30
social	-.32	.13
welcome	-.09	.43
together	.40	.34
kind	-.23	.50
dear	.11	.41
agree	-.30	**.51**
loved	.03	**.51**
relationship	**.73**	.05
liked	.02	.12
loving	.18	.33
boyfriend	**.72**	.10
appreciate	.06	.27
girlfriend	**.66**	.06
helping	-.25	.38
united	-.27	.09

County-level correlations		
Socioeconomic Status (SES)†	-.43 [-.47, -.38]	.14 [.08, .19]
Atherosclerotic Heart Disease	.18 [.13, 23]	-.02 [-.07, .04]

Note. Examination of the eigenvalues and the Scree test revealed a clear two factor structure. Words are ordered in descending frequency of occurrence. Factor scores were imputed through regression (random factors, Thompson's method). Pearson correlations (*r*) are given with 95% confidence intervals in brackets. The 20 words shown account for 89.1% of all word occurrences of the positive relationship dictionary.
† SES index combining standardized high school and college graduation rates, and median income.

Table S6

Top Ten Dictionary Words by Frequency and Their Correlations with Atherosclerotic Heart Disease (AHD)

Anger Dictionary

Top Ten Words	Correlation with AHD Mortality (Pearson r with 95% CIs)	Correlation with AHD Mortality Controlled for Income and Education	Overall Frequency
shit	.12 [.06, .17]	.07 [.02, .13]	2,178,219
fuck	.20 [.15, .25]	.17 [.11, .22]	1,551,388
hate	.23 [.18, .28]	.19 [.13, .24]	1,307,810
damn	.03 [-.02, .09]	-.03 [-.08, .03]	1,252,834
bitch	.13 [.07, .18]	.06 [.01, .12]	864,810
hell	.01 [-.04, .07]	-.05 [-.11, .00]	781,102
fucking	.28 [.23, .33]	.29 [.24, .34]	651,694
mad	.13 [.08, .19]	.09 [.03, .14]	514,694
stupid	.11 [.06, .16]	.06 [.00, .11]	410,894
bitches	.13 [.08, .18]	.09 [.03, .14]	305,033

Negative Relationships Dictionary

Top Ten Words	Correlation with AHD Mortality (Pearson r with 95% CIs)	Correlation with AHD Mortality Controlled for Income and Education	Overall Frequency
hate	.23 [.18, .28]	.19 [.13, .24]	1,307,810
alone	.13 [.08, .18]	.09 [.03, .14]	292,621
jealous	.05 [-.01, .10]	.04 [-.02, .09]	177,374
blame	-.01 [-.07, .04]	-.01 [-.06, .04]	100,930
evil	-.07 [-.13, -.02]	-.07 [-.13, -.02]	94,161
rude	.04 [-.01, .10]	.02 [-.03, .08]	78,552
lonely	.05 [-.01, .10]	.01 [-.05, .06]	70,916
independent	-.04 [-.09, .01]	-.02 [-.08, .03]	39,313
hated	.10 [.05, .15]	.09 [.04, .14]	39,251
ban	-.05 [-.10, .00]	-.02 [-.07, .03]	36,417

Negative Emotions Dictionary

Top Ten Words	Correlation with AHD Mortality (Pearson r with 95% CIs)	Correlation with AHD Mortality Controlled for Income and Education	Overall Frequency
sorry	.04 [-.02, .09]	.04 [-.01, .09]	757,751
mad	.13 [.08, .19]	.09 [.03, .14]	514,694
sad	.00 [-.05, .06]	.00 [-.05, .05]	428,082
scared	.09 [.03, .14]	.03 [-.03, .08]	168,420
pissed	.19 [.14, .24]	.15 [.10, .20]	140,696
crying	.11 [.06, .17]	.09 [.04, .14]	123,994
horrible	.07 [.02, .12]	.08 [.02, .13]	113,522
afraid	.05 [-.01, .10]	.04 [-.02, .09]	104,582
terrible	.03 [-.03, .08]	.06 [.00, .11]	104,195
upset	.10 [.05, .15]	.08 [.02, .13]	93,648

Disengagement Dictionary

Top Ten Words	Correlation with AHD Mortality (Pearson r with 95% CIs)	Correlation with AHD Mortality Controlled for Income and Education	Overall Frequency
tired	.16 [.11, .21]	.10 [.05, .16]	580,979
bored	.18 [.13, .23]	.11 [.05, .16]	411,358
sleepy	-.01 [-.06, .04]	-.10 [-.16, -.05]	157,043
lazy	.04 [-.02, .09]	-.01 [-.06, .04]	138,761
blah	.07 [.02, .12]	.03 [-.02, .09]	110,085
meh	-.02 [-.07, .04]	-.04 [-.09, .01]	53,376
exhausted	.06 [.01, .12]	.09 [.03, .14]	49,955
yawn	-.03 [-.09, .02]	-.03 [-.08, .02]	21,398
distracted	-.06 [-.12, -.01]	-.04 [-.10, .01]	17,998
boredom	.04 [-.01, .10]	.04 [-.02, .09]	17,150

Anxiety Dictionary

Top Ten Words	Correlation with AHD Mortality (Pearson r with 95% CIs)	Correlation with AHD Mortality Controlled for Income and Education	Overall Frequency
crazy	.13 [.08, .18]	.09 [.04, .14]	696,947
pressure	.02 [-.03, .08]	.03 [-.02, .09]	193,805
worry	.05 [-.01, .10]	.02 [-.03, .08]	172,486
scared	.09 [.03, .14]	.03 [-.03, .08]	168,420
awkward	.09 [.04, .15]	.09 [.03, .14]	152,980
scary	-.02 [-.08, .03]	-.02 [-.07, .04]	121,521
fear	-.06 [-.12, -.01]	-.05 [-.10, .00]	120,542
doubt	.09 [.03, .14]	.09 [.03, .14]	115,207
horrible	.07 [.02, .12]	.08 [.02, .13]	113,522
afraid	.05 [-.01, .10]	.04 [-.02, .09]	104,582

Positive Relationships Dictionary

Top Ten Words	Correlation with AHD Mortality (Pearson r with 95% CIs)	Correlation with AHD Mortality Controlled for Income and Education	Overall Frequency
love	.13 [.08, .18]	.08 [.02, .13]	5,375,835
home	.11 [.05, .16]	.10 [.04, .15]	1,907,974
friends	.10 [.05, .15]	.09 [.04, .14]	1,005,756
friend	.05 [.00, .10]	.02 [-.03, .07]	721,639
team	-.07 [-.13, -.02]	-.05 [-.10, .01]	629,910
social	-.08 [-.14, -.03]	-.03 [-.09, .02]	448,731
welcome	-.04 [-.09, .01]	-.02 [-.07, .03]	421,685
together	.00 [-.05, .06]	-.02 [-.07, .04]	398,957
kind	-.09 [-.14, -.03]	-.04 [-.10, .01]	379,906
dear	.02 [-.03, .07]	.02 [-.03, .08]	289,738

Positive Emotion Dictionary

Top Ten Words	Correlation with AHD Mortality (Pearson r with 95% CIs)	Correlation with AHD Mortality Controlled for Income and Education	Overall Frequency
great	-.15 [-.21, -.10]	-.09 [-.15, -.04]	2,375,268
happy	.06 [.01, .12]	.06 [.01, .12]	1,830,533
cool	-.09 [-.14, -.04]	-.06 [-.12, -.01]	972,187
awesome	-.07 [-.12, -.01]	-.02 [-.08, .03]	971,447
amazing	.04 [-.01, .09]	.09 [.04, .15]	715,301
glad	-.07 [-.13, -.02]	-.09 [-.15, -.04]	499,789
excited	.00 [-.06, .05]	.04 [-.01, .09]	495,371
super	-.01 [-.06, .05]	.01 [-.04, .07]	473,677
enjoy	-.07 [-.12, -.01]	-.02 [-.07, .03]	381,689
wonderful	-.05 [-.10, .00]	-.04 [-.09, .02]	204,721

Engagement Dictionary

Top Ten Words	Correlation with AHD Mortality (Pearson r with 95% CIs)	Correlation with AHD Mortality Controlled for Income and Education	Overall Frequency
learn	-.08 [-.13, -.02]	-.05 [-.11, .00]	350,873
interesting	-.17 [-.22, -.12]	-.10 [-.15, -.04]	305,703
awake	.12 [.07, .17]	.11 [.05, .16]	158,400
interested	-.10 [-.15, -.05]	-.05 [-.10, .01]	137,553
alive	.07 [.01, .12]	.06 [.01, .11]	132,898
learning	-.11 [-.16, -.06]	-.07 [-.12, -.02]	118,337
creative	-.10 [-.16, -.05]	-.04 [-.10, .01]	89,367
alert	-.04 [-.09, .01]	-.02 [-.08, .03]	80,982
involved	-.09 [-.14, -.04]	-.05 [-.11, .00]	65,361
careful	-.07 [-.12, -.02]	-.09 [-.14, -.03]	63,719

BIBLIOGRAPHY

Alderson, J. C. (2007). Judging the frequency of English words. *Applied Linguistics*, *28*(3), 383-409.

American Community Survey. (2009). Selected social characteristics in the United States. Retrieved from http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_09_1YR_DP2&prodType=table

American Hospital Association. (2005). AHA annual survey database. *Washington, DC: American Hospital Association*.

Aragonès, E., Piñol, J. L., & Labad, A. (2006). The overdiagnosis of depression in non-depressed patients in primary care. *Family practice*, *23*(3), 363-368.

Atkins, D. C., Rubin, T. N., Steyvers, M., Doeden, M., Baucom, B. R., & Christensen, A. (2012). Topic models: A novel method for modeling couple and family text data. *Journal of Family Psychology*, *26*(5), 816–827. doi:10.1037/a0029607

Auchincloss, A. H., Gebreab, S. Y., Mair, C., & Diez Roux, A. V. (2012). A review of spatial methods in epidemiology, 2000-2010. *Annual Review of Public Health, 33*, 107-122. http://dx.doi.org/10.1146/annurev-publhealth-031811-124655

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders (DSM-5).* American Psychiatric Pub.

Bagroy, S., Kumaraguru, P., & De Choudhury, M. (2017). A Social Media Based Index of Mental Well-Being in College Campuses. In *Proceedings of the 34th Annual ACM Conference on Human Factors in Computing Systems.* Denver, CO.

Bair, M. J., Robinson, R. L., Katon, W., & Kroenke, K. (2003). Depression and pain comorbidity: a literature review. *Archives of Internal Medicine*, *163*(20), 2433-2445.

Basco, M. R., Bostic, J. Q., Davies, D., Rush, A. J., Witte, B., Hendrickse, W., & Barnett, V. (2000). Methods to improve diagnostic accuracy in a community mental health setting. *American Journal of Psychiatry*, *157*(10), 1599-1605.

Behavioral Risk Factors Surveillance Survey. (2009-2010). Annual survey data. Centers for Disease Control and Prevention. Retrieved from www.cdc.gov/brfss/annual_data/annual_data.htm

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289-300.

Benton, A., Mitchell, M., & Hovy, D. (2017). Multi-Task Learning for Mental Health using Social Media Text. In *Proceedings of European Chapter of the Association for Computational Linguistics*. Valencia, Spain.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*, 993–1022. doi:10.1162/jmlr.2003.3.4-5.993

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

Boehm, J. K., & Kubzanky, L. D. (in press). Positive psychological well-being and cardiovascular disease. In W. Kop, L. Katzel, & S. Waldstein (Eds.), *Cardiovascular Behavioral Medicine*. New York: Springer.

Boehm, J. K., & Kubzansky, L. D. (2012). The heart's content: The association between positive psychological well-being and cardiovascular health. *Psychological Bulletin*, *138*, 655-691. http://dx.doi.org/10.1037/a0027448

Boudreaux, E. D., Cagande, C., Kilgannon, H., Kumar, A., & Camargo, C. A. (2006). A prospective study of depression among adult patients in an urban emergency department. *Prim Care Companion J Clin Psychiatry*, *8*(2), 66-70.

Brownstein, J. S., Freifeld, C. C., & Madoff, L. C. (2009). Digital disease detection—harnessing the Web for public health surveillance. *New England Journal of Medicine*, *360*(21), 2153-2157.

Butler, D. (2013). When Google got flu wrong. *Nature*, *494*(7436), 155.

Campbell, R. S., & Pennebaker, J. W. (2003). The secret life of pronouns: Flexibility in writing style and physical health. *Psychological Science*, *14*(1), 60–65.

Cavazos-Rehg, P. A., Krauss, M. J., Sowles, S., Connolly, S., Rosas, C., Bharadwaj, M., & Bierut, L. J. (2016). A content analysis of depression-related tweets. *Computers in Human Behavior*, *54*, 351-357.

Centers for Disease Control and Prevention (CDC). (2010). Underlying cause of death 1999-2010. *CDC WONDER Online Database*. Retrieved from http://wonder.cdc.gov/ucd-icd10.html

Cepoiu, M., McCusker, J., Cole, M. G., Sewitch, M., Belzile, E., & Ciampi, A. (2008). Recognition of depression by non-psychiatric physicians—a systematic literature review and meta-analysis. *Journal of General Internal Medicine*, *23*(1), 25-36.

Chaix, B., Merlo, J., Evans, D., Leal, C., & Havard, S. (2009). Neighbourhoods in eco-epidemiologic research: delimiting personal exposure areas. A response to Riva, Gauvin, Apparicio and Brodeur. *Social Science & Medicine, 69*, 1306-1310. http://dx.doi.org/10.1016/j.socscimed.2009.07.018

Charikar, M. S. (2002). Similarity estimation techniques from rounding algorithms. In *Proceedings of the 34th Annual ACM Symposium on Theory of Computing (ACM*). Quebec, Canada.

Chew, C., & Eysenbach, G. (2010). Pandemics in the age of Twitter: Content analysis of Tweets during the 2009 H1N1 outbreak. *PloS One*, 5, e14118. http://dx.doi.org/10.1371/journal.pone.0014118

Chida, Y., & Steptoe, A. (2008). Positive psychological well-being and mortality: A quantitative review of prospective observational studies. *Psychosomatic Medicine, 70*, 741-756. http://dx.doi.org/10.1097/PSY.0b013e31818105ba

Chida, Y., & Steptoe, A. (2009). The association of anger and hostility with future coronary heart disease: A meta-analytic view of prospective evidence. *Journal of the American College of Cardiology*, *53*, 936-946. http://dx.doi.org/10.1016/j.jacc.2008.11.044

Chung, C., & Pennebaker, J. (2007). The Psychological Functions of Function Words. *Social Communication*, 343–359. doi:10.4324/9780203837702

Cohen, D. A., Farley, T. A., & Mason, K. (2003). Why is poverty unhealthy? Social and physical mediators. *Social Science & Medicine*, *57*, 1631-1641. http://dx.doi.org/10.1016/S0277-9536(03)00015-7

Coppersmith, G., Dredze, M., & Harman, C. (2014a). Quantifying mental health signals in Twitter. *Workshop on Computational Linguistics and Clinical Psychology*, 51-60.

Coppersmith, G., Harman, C., & Dredze, M. (2014b). Measuring Post Traumatic Stress Disorder in Twitter. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media, 579-582*.

Coppersmith, G., Dredze, M., Harman, C., & Hollingshead, K. (2015). From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality (ACL)*. Denver, CO.

Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., & Mitchell, M. (2015). CLPsych 2015 shared task: Depression and PTSD on Twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality (ACL)*. Denver, CO.

Coppersmith, G., Ngo, K., Leary, R., & Wood, A. (2016). Exploratory analysis of social media prior to a suicide attempt. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology (ACL)*. San Diego, CA.

Costa, P. T., Jr., & McCrae, R. R. (1992). Revised NEO Personality Inventory (Neo-PI-R) and NEO Five-Factor Inventory (NEO-FFI): Professional manual. Odessa, FL: Psychological Assessment Resources.

Coyne, J. C., Schwenk, T. L., & Fechner-Bates, S. (1995). Nondetection of depression by primary care physicians reconsidered. *General Hospital Psychiatry*, *17*(1), 3-12.

De Choudhury, M., Counts, S., & Horvitz, E. (2013a). Social media as a measurement tool of depression in populations. In the *Proceedings of the 5th Annual ACM Web Science Conference*. Paris, France.

De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013b). Predicting Depression via Social Media. In *Proceedings of the 7th International Association for the Advancement of Artificial Intelligence Conference on Weblogs and Social Media (ICWSM).* Boston, MA.

De Choudhury, M., Counts, S., Horvitz, E. J., & Hoff, A. (2014). Characterizing and predicting postpartum depression from shared Facebook data. In *Proceedings of the 17th Association for Computing Machinery Conference on Computer Supported Cooperative Work & Social Computing (CSCW).* Baltimore, MD.

De Choudhury, M., Kiciman, E., Dredze, M., Coppersmith, G., & Kumar, M. (2016). Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 33rd Annual Association for Computing Machinery Conference on Human Factors in Computing Systems (ACM)*. San Jose, CA.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391.

Demyttenaere, K., Bruffaerts, R., Posada-Villa, J., Gasquet, I., Kovess, V., Lepine, J., ... & Polidori, G. (2004). Prevalence, severity, and unmet need for treatment of mental disorders in the World Health Organization World Mental Health Surveys. *Journal of the American Medical Association*, *291*(21), 2581-2590.

Diez Roux, A. V., & Mair, C. (2010). Neighborhoods and health. *Annals of the New York Academy of Sciences*, *1186*, 125-145. http://dx.doi.org/10.1111/j.1749-6632.2009.05333.x

Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., & Harshman, R. (1988).

Using latent semantic analysis to improve access to textual information. In *Proceedings of the Special Interest Group on Computer-Human Interaction Conference on Human Factors in Computing Systems (ACM*). Washington, DC.

Edwards, M. & Holtzman, N. S. (2017). A meta-analysis of correlations between depression and first person singular pronoun use. *Journal of Research in Personality*.

Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., … Seligman, M. E. P. (2015). Psychological Language on Twitter Predicts County-Level Heart Disease Mortality. *Psychological Science.* 26(2): 159-169.67

Eysenbach, G. (2009). Infodemiology and infoveillance: Framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. *Journal of Medical Internet Research, 11,* e11. http://dx.doi.org/10.2196/jmir.1157

Facebook: Our mission. (n.d.). Retrieved from https://newsroom.fb.com/company-info/

Fan, H., Yu, W., Zhang, Q., Cao, H., Li, J., Wang, J., ... & Hu, X. (2014). Depression after heart failure and risk of cardiovascular and all-cause mortality: a meta-analysis. *Preventive Medicine*, *63*, 36-42.

Fan, R., Zhao, J., Chen, Y., & Xu, K. (2014). Anger is more influential than joy: Sentiment correlation in Weibo. *PloS One*, *9*(10), e110184.

Fiest, K. M., Jette, N., Quan, H., Germaine-Smith, C. S., Metcalfe, A., Patten, S. B., & Beck, C. A. (2014). Systematic review and assessment of validated case definitions for depression in administrative data. *BMC Psychiatry*, *14*(1), 289.

Ford, E. S., & Capewell, S. (2011). Proportion of the decline in cardiovascular mortality disease due to prevention versus treatments: Public health versus clinical care. *Annual Review of Public Health*, *32*, 5-22. http://dx.doi.org/10.1146/annurev-publhealth-031210-101211

Fox, S., Zickuhr, K., & Smith, A. (2009). Twitter and status updating, Fall 2009. PewResearch Internet Project. Retrieved from http://www.pewinternet.org/2009/10/21/twitter-and-status-updating-fall-2009

Francis, M. E., & Pennebaker, J. W. (1993). LIWC: Linguistic inquiry and word count. *Dallas, TX: Southern Methodist University*.

Francis, M. E., & Pennebaker, J. W. (1992). Putting Stress into Words: The Impact of Writing on Physiological, Absentee, and Self-Reported Emotional Well-Being Measures. *American Journal of Health Promotion*, *6*(4), 280-287

Friedman, H. S., & Kern, M. L. (2014). Personality, well-being, and health. *Annual Review of Psychology*, *65*, 719-742. http://dx.doi.org/10.1146/annurev-psych-010213-115123

Fredrickson, B. L., Mancuso, R. A., Branigan, C., & Tugade, M. M. (2000). The undoing effects of positive emotions. *Motivation and Emotion*, *24*, 237-258. http://dx.doi.org/10.1023/A:1010796329158

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137-144.

Gibbs, S. (2016, February 2). *Whatsapp and Gmail join the 1 billion user club*. Retrieved from https://www.theguardian.com/technology/2016/feb/02/whatsapp-gmail-google-facebook-user-app

Gilbert, E. (2012). Phrases that signal workplace hierarchy. *In Proceedings of the*

Association for Computing Machinery Conference on Computer Supported Cooperative Work (ACM). Seattle, WA.

Gilbody, S., Richards, D., Brealey, S., & Hewitt, C. (2007). Screening for depression in medical settings with the Patient Health Questionnaire (PHQ): a diagnostic meta-analysis. *Journal of General Internal Medicine*, *22*(11), 1596-1602.

Gilbody, S., Sheldon, T., & House, A. (2008). Screening and case-finding instruments for depression: a meta-analysis. *Canadian Medical Association Journal*, *178*(8), 997-1003.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature, 457*, 1012-1014. http://dx.doi.org/10.1038/nature07634

Gkotsis, G., Velupillai, S., Oellrich, A., Dean, H., Liakata, M., & Dutta, R. (2016). Don't Let Notes Be Misunderstood: A Negation Detection Method for Assessing Risk of Suicide in Mental Health Records. In *Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology*, 95-105.

Glaser, B., & Strauss, A. (1967). The discovery of grounded theory. 1967. *Weidenfield & Nicolson, London*, 1-19.

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality Psychology in Europe*, *7*(1), 7-28.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public domain personality measures. *Journal of Research in Personality*, *40*, 84 –

96.

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.

Gottschalk, L. a., & Bechtel, R. (1995). Computerized measurement of the content analysis of natural language for use in biomedical and neuropsychiatric research. *Computer Methods and Programs in Biomedicine*, *47*(2), 123–130. doi:10.1016/0169-2607(95)01645-A

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211–244. doi:10.1037/0033-295X.114.2.211

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis.* http://dx.doi.org/10.1093/pan/mps028

Gross, A., & Murthy, D. (2014). Modeling virtual organizations with Latent Dirichlet Allocation: A case for natural language processing. *Neural Networks*, *58*, 38-49.

Hart, R. P. (1984). *Verbal style and the presidency: A computer-based analysis*. Academic Pr.

Hart, R. P. (1997). *Diction 4.0: The Text-analysis Program: User's Manual*. Scolari.

Hart, R. (2001). Redeveloping Diction: Theoretical Considerations. In *Theory, Method, and Practice in Computer Content Analysis* (pp. 43-60). Westport, CT: Ablex Publishing.

Hegerl, U., Wittmann, M., Arensman, E., Van Audenhove, C., Bouleau, J. H., Van Der

Feltz-Cornelis, C., ... & Meise, U. (2008). The European Alliance against depression-a four-level intervention programme against depression and suicidality. *Suicidologi*, *13*(1).

Henkel, V., Mergl, R., Kohnen, R., Maier, W., Möller, H. J., & Hegerl, U. (2003). Identifying depression in primary care: a comparison of different methods in a prospective cohort study. *BMJ*, *326*(7382), 200-201.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 65-70.

Holsti, O. R., Brody, R. A., & North, R. C. (1964). Measuring Affect and Action in International Reaction Models: Empirical Materials from the 1962 Cuban Missile Cricis. *Journal of Peace Research*, *1*(3-4), 170–189.

Homan, C. M., Lu, N., Tu, X., Lytle, M. C., & Silenzio, V. (2014, February). Social structure and depression in TrevorSpace. In *Proceedings of the 17th Association for Computing Machinery Conference on Computer Supported Cooperative Work & Social Computing*. Baltimore, MD.

Howell, R., Kern, M. L. & Lyubomirsky, S. (2007). Health benefits: Meta-analytically determining the impact of well-being on objective health outcomes. *Health Psychology Review, 1*, 83-136. http://dx.doi.org/10.1080/17437190701492486

Hustey, F. M., & Smith, M. D. (2007). A depression screen and intervention for older ED patients. *The American Journal of Emergency Medicine*, *25*(2), 133-137.

Hwang, J. D., & Hollingshead, K. (2016). Crazy Mad Nutters: The Language of Mental Health. In *Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology*, 52-62.

Iacobelli, F., Gill, A. J., Nowson, S., & Oberlander, J. (2011). Large scale personality classification of bloggers. In *Affective Computing and Intelligent Interaction* (pp. 568-577). Springer Berlin Heidelberg.

Iliev, R., Dehghani, M., & Sagi, E. (2014). Automated text analysis in psychology: methods, applications, and future developments. *Language and Cognition*, 1–26. doi:10.1017/langcog.2014.30

Inkster, B., Stillwell, D., Kosinski, M., & Jones, P. (2016). A decade into Facebook: where is psychiatry in the digital age?. *The Lancet Psychiatry*, *3*(11), 1087-1090.

Kern, M. L., Eichstaedt, J. C., Schwartz, H. A., Dziurzynski, L., Ungar, L. H., Stillwell, D. J., Kosinski, M., Ramones, S. M., & Seligman, M. E. (2014a). The Online Social Self: An Open Vocabulary Approach to Personality. *Assessment*, *21*(2), 158-169.

Kern, M. L., Eichstaedt, J. C., Schwartz, H. A., Park, G., Ungar, L. H., Stillwell, D. J., ... & Seligman, M. E. (2014b). From "Sooo excited!!!" to "So proud": Using language to study development. *Developmental Psychology*, *50*(1), 178.

Kern, M. L., Park, G., Eichstaedt, J. C., Schwartz, H. A., Sap, M., Smith, L. K., & Ungar, L. H. (2016). Gaining insights from social media language: Methodologies and challenges. *Psychological Methods*, 21(4).

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Koretz, D., Merikangas, K. R., ... & Wang, P. S. (2003). The epidemiology of major depressive disorder: results from the National Comorbidity Survey Replication (NCS-R). *Journal of the American Medical Association*, *289*(23), 3095-3105.

Kirmayer, L., Robbins, J., Dworkind, M., & Yaffee, M. J. (1993). Somatization and the recognition of depression and anxiety in primary care. *American Journal of Psychiatry*, *150*(5), 734-741.

Kosinski, M. & Stillwell, D. (2012). mypersonality project. In http://www.mypersonality.org/wiki/.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. In *Proceedings of the National Academy of Sciences of the United States of America*, *110*(15), 5802–5. doi:10.1073/pnas.1218772110

Kosinski, M. (2014). *'Measurement and prediction of individual and group differences in the digital environment* (Doctoral dissertation, Ph. D. dissertation, Dept. Psychol., Cambridge Univ., Cambridge, UK).

Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V., & Stillwell, D. (2015). Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist*, *70*(6), 543.

Kramer, A. D. (2010). An unobtrusive behavioral model of gross national happiness. In *Proceedings of the Special Interest Group on Computer-Human Interaction Conference on Human Factors in Computing Systems (ACM*). Atlanta, GA.

Labarthe, D. R. (2010). *Epidimiology and prevention of cardiovascular disease: A global challenge*. Sudbury, MA: Jones and Bartlett Publishers.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240. doi:10.1037/0033-295X.104.2.211

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2-3), 259-284.

Lasswell, H. D., & Kaplan, A. (1950). *Power and society: A framework for political inquiry*. Transaction Publishers.

Lasswell, H. D., & Namenwirth, J. Z. (1969). The Lasswell value dictionary. *New Haven*.

Lawless, N. M., & Lucas, R. E. (2011). Predictors of regional well-being: A county level analysis. *Social Indicators Research*, *101*(3), 341-357.

Lett, H. S., Blumenthal, J. A., Babyak, M. A., Sherwood, A., Strauman, T., Robins, C., & Newman, M. F. (2004). Depression as a risk factor for coronary artery disease: Evidence, mechanisms, and treatment. *Psychosomatic Medicine*, *66*, 305-315. http://dx.doi.org/10.1097/01.psy.0000126207.43307.c0

Leyland, A. H. (2005). Socioeconomic gradients in the prevalence of cardiovascular disease in Scotland: The roles of composition and context. *Journal of Epidemiology and Community Health*, *59*, 799-803. http://dx.doi.org/10.1136/jech.2005.034017

Liu, L., Preotiuc-Pietro, D., Samani, Z. R., Moghaddam, M. E., & Ungar, L. H. (2016). Analyzing Personality through Social Media Profile Picture Choice. In *Proceedings of the 10th International Association for the Advancement of Artifical Intelligence Conference on Web and Social Media (ICWSM)*. Cologne, Germany.

Lloyd-Jones, D. M., Hong, Y., Labarthe, D., Mozaffarian, D., Appel, L. J., Van Horn, L., …, & Rosamond, W. D. (2010). Defining and setting national goals for cardiovascular health promotion and disease reduction the American Heart Association's strategic impact goal through 2020 and beyond. *Circulation*, *121*, 585-613. http://dx.doi.org/10.1161/CIRCULATIONAHA.109.192703

Lochner, K. A., Kawachi, I., Brennan, R. T., & Buka, S. L. (2003). Social capital and neighborhood mortality rates in Chicago. *Social Science & Medicine, 56*, 1797-1805. http://dx.doi.org/10.1016/S0277-9536(02)00177-6

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity.

Martindale, C. (1973). An experimental simulation of literary change. *Journal of Personality and Social Psychology*, *25*(3), 319–326. doi:10.1037/h0034238

Martindale, C. (1975). *The romantic progression: the psychology of literary history*. Halsted Press.

McCallum, A. K. (2002). Mallet: A machine learning for language toolkit.

McClelland, D. C., Davis, W., Wanner, E., Kalin, R., Mcclelland, D. C., Davis, W., & Wanner, E. (1966). A Cross-Cultural Study of Folk-Tale Content and Drinking *. *Sociometry*, *29*(4), 308–333.

McKee, R. (2013). Ethical issues in using social media for health and health care research. *Health Policy*, *110*(2), 298-301.

McKelvey, K., DiGrazia, J., & Rojas, F. (2014). Twitter publics: How online political communities signaled electoral outcomes in the 2010 US house election. *Information, Communication & Society*, *17*(4), 436-450.

Meehl, P. E. (1954). Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. Minneapolis, MN, US: University of Minnesota Press.

Mehl, M. R. (2006). Quantitative text analysis. *Handbook of Multimethod Measurement in Psychology*, 141-156.

Menezes, A. R., Lavie, C. J., Milani, R. V., O'Keefe, J., & Lavie, T. J. (2011). Psychological risk factors and cardiovascular disease: Is it all in your head? *Postgraduate Medication, 123*, 165-176. http://dx.doi.org/10.3810/pgm.2011.09.2472

Mergenthaler, E., & Bucci, W. (1999). Linking verbal and non-verbal representations: computer analysis of referential activity. *The British Journal of Medical Psychology*, *72 (Pt 3)*, 339–354. doi:10.1348/000711299160040

Miller, P. R., Dasher, R., Collins, R., Griffiths, P., & Brown, F. (2001). Inpatient diagnostic assessments: 1. Accuracy of structured vs. unstructured interviews. *Psychiatry Research*, *105*(3), 255-264.

Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J. P., & Rosenquist, J. N. (2011). Understanding the demographics of Twitter users. In the *Proceedings of the 5th International Association for the Advancement of Artificial Intelligence Conference on Weblogs and Social Media (ICWSM 2011)*. Barcelona, Spain.

Mitchell, A. J., & Coyne, J. C. (2007). Do ultra-short screening instruments accurately detect depression in primary care? *British Journal of General Practice*, *57*(535), 144-151.

Mitchell, A. J., Vaze, A., & Rao, S. (2009). Clinical diagnosis of depression in primary care: a meta-analysis. *The Lancet*, *374*(9690), 609-619.

Mitchell, A. J., Rao, S., & Vaze, A. (2011). International comparison of clinicians' ability to identify depression in primary care: meta-analysis and meta-regression of predictors. *British Journal of General Practice*, *61*(583), e72-e80.

Mobile Fact Sheet. (2017, January 12). Retrieved from http://www.pewresearch.org/about/use-policy/

Mohr, D. C., Zhang, M., & Schueller, S. (2017). Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning. *Annual Review of Clinical Psychology*, *13*(1).

Mojtabai, R. (2013). Clinician-identified depression in community settings: concordance with structured-interview diagnoses. *Psychotherapy and Psychosomatics*, *82*(3), 161-169.

Morgan, C. D., & Murray, H. A. (1935). A Method for Investigating Fantasies: The Thematic Apperception Test. *Archives of Neurology and Psychiatry*, *34*(2), 289–306.

Mowery, D. L., Bryan, C., & Conway, M. (2015). Towards Developing an Annotation Scheme for Depressive Disorder Symptoms: A Preliminary Study using Twitter Data. In *Proceeding of 2nd Workshop on Computational Linguistics and Clinical Psychology-From Linguistic Signal to Clinical Reality*. Denver, CO.

Mowery, D., Bryan, C., & Conway, M. (2017). Feature studies to inform the classification of depressive symptoms from Twitter data for population health. *arXiv preprint arXiv:1701.08229*.

Mulrow, C. D., Williams, J. W., Gerety, M. B., Ramirez, G., Montiel, O. M., & Kerber, C. (1995). Case-finding instruments for depression in primary care settings. *Annals of Internal Medicine*, *122*(12), 913-921.

Murray, H. A. (1938). *Explorations in personality*. Oxford, England: Oxford Univ. Press.

Murray, H. A. (1943). Thematic apperception test. *Neuendorf*, 2002

Nadeem, M. (2016). Identifying depression on Twitter. *arXiv preprint arXiv:1607.07384*.

National Institute of Mental Health (2015). Major depression among adults. Retrieved from https://www.nimh.nih.gov/health/statistics/prevalence/major-depression-among-adults.shtml

Nease, D. E., & Malouin, J. M. (2003). Depression screening: a practical strategy. *Journal of Family Practice*, *52*(2), 118-126.

Newman, M. L., Groom, C. J., Handelman, L. D., & Pennebaker, J. W. (2008). Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, *45*(3), 211-236.

Newman, M. G., Szkodny, L. E., Llera, S. J., & Przeworski, A. (2011). A review of technology-assisted self-help and minimal contact therapies for anxiety and depression: is human contact necessary for therapeutic efficacy?. *Clinical Psychology Review*, *31*(1), 89-103.

Noyes, K., Liu, H., Lyness, J. M., & Friedman, B. (2011). Medicare beneficiaries with depression: comparing diagnoses in claims data with the results of screening. *Psychiatric Services*, *62*(10), 1159-1166.

O'Malley, K. J., Cook, K. F., Price, M. D., Wildes, K. R., Hurdle, J. F., & Ashton, C. M. (2005). Measuring diagnoses: ICD code accuracy. *Health Services Research*, *40*(5p2), 1620-1639.

Osgood, S., Suci, G. J., and Tannenbaum, P.H. (1957). The measurement of meaning. *Urbana: University of Illinois Press*.

Osgood, C. E. (1963). On understanding and creating sentences. *American Psychologist*, *18*(12), 735.

Oxman, T. E., Dietrich, A. J., Williams, J. W., & Kroenke, K. (2002). A three-component model for reengineering systems for the treatment of depression in primary care. *Psychosomatics*, *43*(6), 441-450.

Padrez, K. A., Ungar, L., Schwartz, H. A., Smith, R. J., Hill, S., Antanavicius, T., ... & Merchant, R. M. (2015). Linking social media and medical record data: a study of adults presenting to an academic, urban emergency department. *BMJ Quality & Safety*.

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Stillwell, D. J., Ungar, L. H., … Kosinski, M. (2014). Automatic Personality Assessment Through Social Media Language. *Journal of Personality and Social Psychology*, *108(6*), 934-952.

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., ... & Seligman, M. E. (2015). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, *108*(6), 934.

Park, G., Yaden, D. B., Schwartz, H. A., Kern, M. L., Eichstaedt, J. C., Kosinski, M., ... & Seligman, M. E. (2016). Women are Warmer but No Less Assertive than Men: Gender and Language on Facebook. *PloS ONE*, *11*(5), e0155885.

Passik, S. D., Dugan, W., McDonald, M. V., Rosenfeld, B., Theobald, D. E., & Edgerton, S. (1998). Oncologists' recognition of depression in their patients with cancer. *Journal of Clinical Oncology*, *16*(4), 1594-1600.

Paul, M. J., & Dredze, M. (2011a). A model for mining public health topics from Twitter. *Technical Report*. Johns Hopkins University.

Paul, M. J., & Dredze, M. (2011b). You are what you tweet: Analyzing Twitter for public health. In the proceedings of the *5th International Association for the Advancement of Artificial Intelligence Conference on Weblogs and Social Media*

(ICWSM 2011). Barcelona, Spain.

Pedersen, T. (2015). Screening Twitter users for depression and PTSD with lexical decision lists. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Denver, CO.

Pennebaker, J. W. (1997a). Opening up: The healing power of emotional expression. *New York: Guilford*.

Pennebaker, J. W. (1997b). Writing About Emotional Experiences as a Therapeutic Process. *Psychological Science, 8*(3), 162–166. doi:10.1111/j.1467-9280.1997.tb00403.x

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic Inquiry and Word Count (LIWC): A computerized text analysis program. *Mahwah (NJ), 7*.

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (2007). *The development and psychometric properties of LIWC2007*. Austin, TX: LIWC.net.

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology, 54,* 547-577. http://dx.doi.org/10.1146/annurev.psych.54.101601.145041

Pennebaker, J. W. (2011). The secret life of pronouns. *New Scientist, 211*(2828), 42-45.

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (2015). LIWC.

Perruche, F., Elie, C., d'Ussel, M., Ray, P., Thys, F., Bleichner, G., ... & Le Joubioux, E. (2011). Anxiety and depression are unrecognized in emergency patients admitted to the observation care unit. *Emergency Medicine Journal*, emj-2009.

Peterson, C., Luborsky, L., & Seligman, M. E. P. (1983). Attributions and Depressive

Mood Shifts: A Case Study Using the Symptom-Context Method. *Journal of Abnormal Psychology*, *92*(1), 96–103. doi:10.1080/01425690701737481

Peterson, C., & Semmel, A. (1982). The Attributional Style Questionnaire1. *Cognitive Therapy and Research*, *6*(3).

Pierce, J. (1980). *An introduction to information theory: Symbols, signals & noise* (2nd, rev. ed.). New York: Dover Publications.

Pierce, J. R., & Denison, A. V. (2010). Accuracy of death certificates and the implications for studying disease burdens. In Preedy V. R., & R. R. Watson (eds.), *Handbook of Disease Burdens and Quality of Life Measures* (pp. 329-344). New York: Springer.

Potts, C. (2011).*happyfuntokenizer*(Version 10). [Computer software]. Retrieved from http://sentiment.christopherpotts.net/code-data/happyfuntokenizing.py

Preotiuc-Pietro, D., Eichstaedt, J., Park, G., Sap, M., Smith, L., Tobolsky, V., ... & Ungar, L. (2015). The role of personality, age and gender in tweeting about mental illnesses. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology*. Denver, CO.

Pressman, S. D., & Cohen, S. (2005). Does positive affect influence health? *Psychological Bulletin*, *131*, 925-971. http://dx.doi.org/10.1037/0033-2909.131.6.925

Quan, H., Li, B., Duncan Saunders, L., Parsons, G. A., Nilsson, C. I., Alibhai, A., & Ghali, W. A. (2008). Assessing validity of ICD-9-CM and ICD-10 administrative data in recording clinical conditions in a unique dually coded database. *Health Services Research*, *43*(4), 1424-1441.

Quincey, E. D., & Kostkova, P. (2009). Early warning and outbreak detection using social networking websites: The Potential of twitter. Paper presented at the *2nd International Conference on Software Testing.* Istanbul, Turkey.

Reece, A. G., Reagan, A. J., Lix, K. L., Dodds, P. S., Danforth, C. M., & Langer, E. J. (2016). Forecasting the onset and course of mental illness with Twitter data. *arXiv preprint arXiv:1608.07740*.

Resnik, P., Armstrong, W., Claudino, L., Nguyen, T., Nguyen, V. A., & Boyd-Graber, J. (2015). Beyond LDA: exploring supervised topic modeling for depression-related language in Twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Denver, CO.

Riva, M., Gauvin, L., & Barnett, T. A. (2007). Toward the next generation of research into small area effects on health: A synthesis of multilevel investigations published since July 1998. *Journal of Epidemiology and Community Health, 61*, 853-861. http://dx.doi.org/10.1136/jech.2006.050740

Roest, A. M., Martens, E. J., de Jonge, P., & Denollet, J. (2010). Anxiety and risk of incident coronary heart disease. *Journal of the American College of Cardiology, 56*, 38-46. http://dx.doi.org/10.1016/j.jacc.2010.03.034

Rorschach, H. (1942). Psychodiagnostics. Oxford, England.

Rosenthal, R., & DiMatteo, M. R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology*, *52,* 59-82. http://dx.doi.org/10.1146/annurev.psych.52.1.59

Rost, K., Zhang, M., Fortney, J., Smith, J., Coyne, J., & Smith, G. R. (1998). Persistently poor outcomes of undetected major depression in primary care. *General Hospital Psychiatry, 20*(1), 12-20.

Rost, M., Barkhuus, L., Cramer, H., & Brown, B. (2013, February). Representation and communication: Challenges in interpreting large social media datasets. Paper presented at *the 16th Association for Computing Machinery Conference on Computer Supported Cooperative Work and Social Computing*. San Antonio, TX.

Rowe, J. W., & Kahn, R. L. (1987). Human aging: Usual and successful. *Science, 237,* 143-149. http://dx.doi.org/10.1126/science.3299702

Ruddick, G. (2016). Admiral to price car insurance based on Facebook posts. *The Guardian*. Retrieved from https://www.theguardian.com/technology/2016/nov/02/admiral-to-price-car-insurance-based-on-facebook-posts

Rugulies, R. (2002). Depression as a predictor for coronary heart disease: A review and meta-analysis. *American Journal of Preventive Medicine, 23,* 51-61. http://dx.doi.org/10.1016/S0749-3797(02)00439-7

Rush, J. A. (1993). *Depression Guideline Panel. Depression in primary care: Volume 2: Treatment of major depression*. Clinical Practice Guideline, Number 5.

Salathé, M., Freifeld, C. C., Mekaru, S. R., Tomasulo, A. F., & Brownstein, J. S. (2013). Influenza A (H7N9) and the importance of digital epidemiology. *New England Journal of Medicine, 369,* 401-404. http://dx.doi.org/10.1056/NEJMp1307752

Sap, M., Park, G., Eichstaedt, J. C., Kern, M. L., Stillwell, D., Kosinski, M., ... & Schwartz, H. A. (2014). Developing age and gender predictive lexica over social media. In *Proceedings of the 2014 Conference on Empirical Methods in Natural*

Language Processing. Doha, Qatar.

Schulberg, H. C., Saul, M., McClelland, M., Ganguli, M., Christy, W., & Frank, R. (1985). Assessing depression in primary medical and psychiatric practices. *Archives of General Psychiatry*, *42*(12), 1164-1170.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Lucas, R. E., Agrawal, M., Park, G. J., Lakshmikanth, S. K., Jha, S., Seligman, M. E. P., & Ungar, L. H. (2013). Characterizing Geographic Variation in Well-Being using Tweets. *In Seventh International AAAI Conference on Weblogs and Social Media (ICWSM)*. Boston, MA.

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

Schwartz, H. A., Eichstaedt, J. C., Dziurzynski, L., Kern, M. L., Blanco, E., Kosinski, M., ... & Ungar, L. H. (2013c). Toward Personality Insights from Language Exploration in Social Media. In *AAAI Spring Symposium: Analyzing Microtext*.

Schwartz, H. A., Eichstaedt, J., Kern, M. L., Park, G., Sap, M., Stillwell, D., ... & Ungar, L. (2014). Towards assessing changes in degree of depression through facebook. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Baltimore, MD.

Schwartz, H. A., & Ungar, L. H. (2015). Data-driven content analysis of social media: a systematic overview of automated methods. *The ANNALS of the American Academy of Political and Social Science, 659*(1), 78-94.

Seabrook, E. M., Kern, M. L., & Rickard, N. S. (2016). Social networking sites, depression, and anxiety: a systematic review. *JMIR Mental Health*, *3*(4), e50.

Seifter, A., Schwarzwalder, A., Geis, K., & Aucott, J. (2010). The utility of “Google Trends” for epidemiological research: Lyme disease as an example. *Geospatial Health*, *4,* 135-137.

Seyfried, L., Hanauer, D. A., Nease, D., Albeiruti, R., Kavanagh, J., & Kales, H. C. (2009). Enhanced identification of eligibility for depression research using an electronic medical record search engine. *International Journal of Medical Informatics*, *78*(12), e13-e18.

Simon, G. E., VonKorff, M., Piccinelli, M., Fullerton, C., & Ormel, J. (1999). An international study of the relation between somatic symptoms and depression. *New England Journal of Medicine*, *1999*(341), 1329-1335.

Smith, C. P. (Ed.). (1992). *Motivation and personality: Handbook of thematic content analysis*. Cambridge University Press

Solberg, L. I., Engebretson, K. I., Sperl-Hillen, J. M., Hroscikoski, M. C., & O'connor, P. J. (2006). Are claims data accurate enough to identify patients for performance measures or quality improvement? The case of diabetes, heart disease, and depression. *American Journal of Medical Quality*, *21*(4), 238-245.

Sorg, S., Vögele, C., Furka, N., & Meyer, A. H. (2012). Perseverative thinking in depression and anxiety. *Frontiers in Psychology*, *3*.

Spertus, E., Sahami, M., & Buyukkokten, O. (2005, August). Evaluating similarity measures: a large-scale study in the orkut social network. In *Proceedings of the eleventh ACM SIGKDD International Conference on Knowledge Discovery in*

Data Mining (pp. 678-684). ACM.

Spettell, C. M., Wall, T. C., Allison, J., Calhoun, J., Kobylinski, R., Fargason, R., & Kiefe, C. I. (2003). Identifying Physician-Recognized Depression from Administrative Data: Consequences for Quality Measurement. *Health Services Research*, *38*(4), 1081-1102.

St Louis, C., & Zorlu, G. (2012, May). Can Twitter predict disease outbreaks? *British Medical Journal, 344*, e2353. http://dx.doi.org/10.1136/bmj.e2353

Stillwell, D. J., & Kosinski, M. (2004). mypersonality project: Example of successful utilization of online social networks for large-scale social research. *American Psychologist*, *59*(2), 93-104.

Stone, P. J., Bales, R. F., Namenwirth, J. Z., & Ogilvie, D. M. (1962). The General Inquirer: A Computer System for Content Analysis and Retrieval Based on the Sentence as Unit of Information. *Computers in Behavioral Science, 7*(4), 484–498. doi:10.1080/01425690701737481

Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1966). *The general inquirer: A computer approach to content analysis.* Cambridge, MA: MIT press.

Tay, L., Tan, K., Diener, E., & Gonzalez, E. (2013). Social relations, health behaviors, and health outcomes: A survey and synthesis. *Applied Psychology: Health and Well-Being, 5*, 28-78.

Trinh, N. H. T., Youn, S. J., Sousa, J., Regan, S., Bedoya, C. A., Chang, T. E., ... & Yeung, A. (2011). Using electronic medical records to determine the diagnosis of clinical depression. *International Journal of Medical Informatics*, *80*(7), 533-540.

Tsugawa, S., Kikuchi, Y., Kishino, F., Nakajima, K., Itoh, Y., & Ohsaki, H. (2015).

Recognizing depression from twitter activity. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. Seoul, Korea.

U.S. Census Bureau. (2010). Profile of general population and housing characteristics. Retrieved from http://factfinder2.census.gov/

Vuorilehto, M., Melartin, T., & Isometsä, E. (2005). Depressive disorders in primary care: recurrent, chronic, and co-morbid. *Psychological Medicine*, *35*(05), 673-682.

Wang, P. S., Lane, M., Olfson, M., Pincus, H. A., Wells, K. B., & Kessler, R. C. (2005). Twelve-month use of mental health services in the United States: results from the National Comorbidity Survey Replication. *Archives of General Psychiatry*, *62*(6), 629-640.

Weber, R. P. (1984). Computer-aided content analysis: A short primer. *Qualitative Sociology*, *7*(1-2), 126-147.

Weber, R.P. (Ed.). (1990). *Basic content analysis*. Sage.

Wilmot, M. P., DeYoung, C. G., Stillwell, D., & Kosinski, M. (2015). Self-Monitoring and the Metatraits.

Wolfe, M. B., & Goldman, S. R. (2003). Use of latent semantic analysis for predicting psychological phenomena: Two issues and proposed solutions. *Behavior Research Methods*, Instruments, & Computers, 35(1), 22-31.

World Health Organization (2011). *Global status report on noncommunicable diseases 2010*. Retrieved from www.who.int/nmh/publications/ncd_report2010/en/

Word Net. (n.d.). Retrieved May 1, 2015, from http://wordnetweb.princeton.edu/

Yang, A. C., Huang, N. E., Peng, C. K., & Tsai, S. J. (2010). Do seasons have an influence on the incidence of depression? The use of an internet search engine query data as a proxy of human affect. *PloS One*, *5*(10), e13728.

Yarkoni, T. (2010). Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality*,*44*(3), 363-373.

Yom-Tov, E., White, R. W., & Horvitz, E. (2014). Seeking insights about cycling mood disorders via anonymized search logs. *Journal of Medical Internet Research*, *16*(2), e65.

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.

Yu, Y., & Wang, X. (2015). World Cup 2014 in the Twitter World: A big data analysis of sentiments in US sports fans' tweets. *Computers in Human Behavior*, *48*, 392-400.

Zimerle, B. K. (2010). *Visible Language: Inventions of Writing in the Ancient East and Beyond*. Chicago: Oriental Institute Museum Publications.

Zimmermann, J., Brockmeyer, T., Hunn, M., Schauenburg, H., & Wolf, M. (2016). First-person Pronoun Use in Spoken Language as a Predictor of Future Depressive Symptoms: Preliminary Evidence from a Clinical Sample of Depressed Patients. *Clinical Psychology & Psychotherapy*.

Zivin, K., Yosef, M., Miller, E. M., Valenstein, M., Duffy, S., Kales, H. C., ... & Kim, H. M. (2015). Associations between depression and all-cause and cause-specific risk of death: a retrospective cohort study in the Veterans Health Administration. *Journal of Psychosomatic Research*, *78*(4), 324-331.