Gaining Insights from Social Media Language: Methodologies and Challenges

Margaret L. Kern[1], Gregory Park[2], Johannes C. Eichstaedt[2], H. Andrew Schwartz[2,3], Maarten Sap[2], Laura K. Smith[2], and Lyle H. Ungar[2]
[1] The University of Melbourne, [2] University of Pennsylvania, [3] Stony Brook University

Author Note

Margaret L. Kern, Melbourne Graduate School of Education, The University of Melbourne, Australia; Gregory Park, Department of Psychology, University of Pennsylvania; Johannes C. Eichstaedt, Department of Psychology, University of Pennsylvania; H. Andrew Schwartz, Computer & Information Science, University of Pennsylvania and Computer Science, Stony Brook University; Maarten Sap, Department of Psychology, University of Pennsylvania; Laura K. Smith, Department of Psychology, University of Pennsylvania; Lyle H. Ungar, Computer & Information Science, University of Pennsylvania.

Support for this publication was provided by the Templeton Religion Trust, grant #TRT0048.

Correspondence concerning this article should be addressed to Margaret L. Kern, Melbourne Graduate School of Education, The University of Melbourne, 100 Leicester Street, Level 2, Parkville, VIC 3010, Australia. Email: margaret.kern@unimelb.edu.au

Main text word count: 11,037 (main text + footnotes, abstract = 135 words)

Final accepted version, April 2016, Psychological Methods. This paper is not the copy of record and may not exactly replicate the authoritative document published in the journal. The final article is available, at http://dx.doi.org/10.1037/met0000091

Abstract

Language data available through social media provide opportunities to study people at an unprecedented scale. However, little guidance is available to psychologists who want to enter this area of research. Drawing on tools and techniques developed in natural language processing, we first introduce psychologists to social media language research, identifying descriptive and predictive analyses that language data allow. Second, we describe how raw language data can be accessed and quantified for inclusion in subsequent analyses, exploring personality as expressed on Facebook to illustrate. Third, we highlight challenges and issues to be considered, including accessing and processing the data, interpreting effects, and ethical issues. Social media has become a valuable part of social life, and there is much we can learn by bringing together the tools of computer science with the theories and insights of psychology.

Keywords: social media, linguistic analysis, interdisciplinary collaboration, online behavior, computational social science

Gaining Insights from Social Media Language: Methodologies and Challenges

The past decade has demonstrated an obsession with data — lots of data. Technological advances make it possible to collect and analyze data at levels never before imagined. Social media provides an active laboratory, far removed from the contrived small-scale experiments that have long dominated psychology. Billions of words, pictures, and behaviors are recorded each day by individuals all around the world. Social media platforms such as Facebook and Twitter have enabled the collection of massive amounts of linguistic information, which reveal individual characteristics and social behaviors (Anderson, Fagan, Woodnutt, & Chamorro-Premuzic, 2012; Gill, 2004; Kern et al., 2014a).

For social scientists that work for months or years to collect data from a few hundred people, the idea of the massive amounts of data available through social media can be both tantalizing and terrifying. Traditional analytic techniques taught in introductory statistics and research method courses are inadequate for dealing with the complexities of such data, and leave little guidance as to how to even begin to approach social media data. Large-scale language analysis is of wide interest, and this paper aims to facilitate an overview and introduction for novel and intermediate researchers.

We first introduce psychologists to research on social media language. Second, we describe how the raw language data can be acquired, processed, and quantified for inclusion in subsequent statistical analyses. We describe steps for accessing and preparing social media language data for statistical analysis, including choosing and obtaining an appropriate dataset, converting the data into a workable format, and top down and bottom up approaches to quantifying information. Depending upon what data are available and the research questions of interest, this process offers many choices. We provide some guidance, and point to additional resources. Finally, despite the appeal of big data, there is little guidance available on problematic issues arising from social media language data. We highlight several aspects here, with recommendations for analysts.

An Introduction to Social Media Language Research

Understanding associations between language and thought have long been an important and vibrant area of research within psychology. However, studying language can require time-intensive qualitative approaches, often with only a handful of respondents. Computational linguistics offers techniques to study language at scale, requiring considerably less time and resources. No longer constrained by results based on small (and often unrepresentative) samples of people, language offers many opportunities to directly study people's thoughts and emotions. Yet as data move from gigabytes to terabytes to petabytes, finding an interpretable signal becomes a process of hunting for a needle in a hay field. Theories are needed to interpret data, and psychologists have developed such theories across hundreds of years. Further benefit can come from collaboration with experts from multiple fields, including quantitative psychologists, statisticians, methodologists, economists, political scientists, health professionals, and educators.

Big Data

The obsession with data has grown exponentially over the past century. As early as the 1940s, discussion began around the "information explosion" and rapid growth of data (see Press, 2013 for a brief history). The term "big data" was first used in 1997: "Visualization provides an interesting challenge for computer systems: data sets are generally quite large, taxing the capacities of main memory, local disk, and even remote disk. We call this the problem of *big data*" (Cox and Ellsworth, 1997, p. 235).

There is now considerable work defining the term "big data" (e.g., Borgman, 2015), such that the term itself is increasingly viewed as unhelpful. From the computer science perspective, social media data can be big in terms of the number of features within each observation (e.g., the number of different words that people use), the number of observations (e.g., the number of people, tweets or Facebook posts), or the total amount of disk space needed to store it. Each of these types of 'bigness' presents different challenges, ranging from appropriate statistical methods to computer software and hardware designed for data-intensive computation. Through aggregation, big data potentially offers data-driven insights – a process quite different from classical hypothesis-driven research.

Our use of "big data" refers to data that allows research to be conducted at an unprecedented scale (Meyer & Schroeder, 2014). Through platforms such as Facebook and Twitter, emails, text messages, and forums, people share a considerable amount of linguistic information. A growing amount of data is also available in the form of explicit behaviors (e.g., "likes", survey answers, pages browsed), unobtrusively monitored behaviors (e.g., steps per day, location, time spent online), and images.

Our primary focus is on data that do not fit in an Excel or SPSS file due to their size and complexity, and thus require different handling than the methods typically used by psychologists. Excellent longitudinal datasets have collected substantial information on hundreds of people's lives, which can be used to study their individual life trajectories (e.g., Block, 1993; Booth et al., 2014; Friedman & Martin, 2011; Hampson et al., 2013; Vaillant, 2012). Epidemiological surveys collect questionnaire data on many people at single time points. In contrast, we focus on data that comes from online behaviors, such as posts on social media, web browsing, shopping, gaming, using mobile applications (Rozenfeld, 2014); they represents digital traces of people as they go about their lives. As such, the data tends to be unstructured, appear in multiple unknown contexts, and be collected with no guiding questions or theory. Further, we specifically focus on language data that comes from social media and related online content. Many of the techniques also apply to other types of linguistic data, which we refer to as "smaller data".

To address many psychologically relevant questions, only the initial processing of data requires so much computing power that it cannot easily be performed on a desktop computer. Consider analyzing how language and personality vary across U.S. counties. Starting with over a billion tweets, each tweet needs to be geolocated (i.e., determine which county it comes from) and tokenized (i.e., broken up into separate "words"). This first pass could take weeks on a single processor computer and so is often done on a powerful cluster (i.e., many computers that are connected together to provide significantly greater processing power). This results in a smaller dataset that contains the counts of how often each word is used in each county. A

smaller dataset indicating the 25,000 most frequent words that occur in 2,000 counties has *only* 50 million entries in it, and can be analyzed on a single server machine (with more than 100 GB of RAM memory and some patience, as this still takes longer than most psychological study analyses).

Types of Analyses

The amount and type of data available impacts the types of analyses that can be performed. Corresponding with the typical psychological analytic tasks of descriptive and inferential studies, language data can be used descriptively to gain insights about individuals and communities and inferentially to make predictions.

Descriptive Studies. Social media data can be used for secondary analysis of existing data. Massive amounts of data are stored by every application, social media platform, email host, website, etc., often with more information-rich text across long time periods than any cross-sectional or short-term questionnaire-based study envisioned in psychology. Starting with particular research questions, appropriate datasets can be selected and analyzed, and new insights derived. In turn, insights and discoveries can be tested in other samples with other designs, cumulatively building the science over time. Hundreds of dissertations could be written with little need to collect additional data. Like the records of anthropology, an entire era of human history has been stored, awaiting further exploration.

Social media language can provide insight into personality and individual characteristics at a more fine-grained level, with attention to how they unfold in the real world. For example, in our analyses of Big Five personality characteristics (Kern et al., 2014a; Schwartz et al., 2013b), some words were intuitive (e.g., individuals high in extraversion used words such as “party” and “great night”), whereas other words revealed more surprising insights (e.g., individuals high in conscientiousness used words such as “relaxing”, “weekend”, and “workout”). Language can extend our understanding of the affective, cognitive, and behavioral processes that characterize these and other constructs. Further, if a survey is given while status updates from social media are collected for the respondent, the words that distinguish different traits or characteristics can be determined, informing our understanding of the construct itself.

Prediction. Language models can potentially be used in various practical ways at both the individual and community levels. There is considerable interest in using social media for real time monitoring of problematic behaviors. Facebook now has a way that people can alert possible suicidal tendencies in a friend, with the hope that timely intervention can occur (https://www.facebook.com/TheSuicideWatchProject). Numerous mobile applications are being developed that allow people to monitor their mood, physiological metrics, sleep, and other behaviors, with automatic alerts if they get off track.

A growing literature stemming from sociology focuses on community characteristics that predict health and well-being. Studies suggest that neighborhood factors, such as food and recreational resources, the built environment, quality of housing, disadvantage, deprivation, feelings of safety, norms, and social connections impact health outcomes in individuals (Chaix, Linstrom, Rosvall, & Merlo, 2008; Diez-Roux & Mair, 2010). However, a constant challenge is how to assess the neighborhood environment. Public health and epidemiological methods

often rely on costly questionnaires. Social media language data provide opportunities for identifying contextual aspects that influence individual and community outcomes, in a much more cost and resource efficient manner. For example, geo-tagged search queries from consented mobile phones predicted health care utilization and duration of hospital stays (Yang, White, & Horvitz, 2013), and search logs identified adverse drug reactions (White, Harpaz, Shah, DuMouchel, & Horvitz, 2014).

Working With Social Media Language Data

With this background on social media language data and potential uses of it, we next turn to methods for working with such data, highlighting key steps for accessing and preparing information.

Illustrative Example: Personality on Facebook

To illustrate the process of working with social media language data, we provide an example based on our own work. The World Well-Being Project (WWBP) is an interdisciplinary collaboration of computer scientists and psychologists. The project began in 2011, with an initial goal of using social media to unobtrusively measure well-being. Whereas social media had been mined heavily for sentiment (i.e., positive and negative emotion), we aimed to capture more holistic elements of well-being, such as social relationships, a sense of meaning or purpose in life, accomplishment, and engagement with life. Following the precedent of Pennebaker and colleagues (2003), we began by manually creating lexica (i.e., dictionaries or lists of words) that theoretically are relevant to different well-being domains. However, when we connected with computer scientists, we recognized that our methods ignored important complexities of linguistic data and that automated analyses could be far more powerful.

Interdisciplinary work is rewarding but challenging. The first year of the collaboration involved learning how to communicate across the two disciplines. Our focus shifted from measuring well-being to understanding individual characteristics that are expressed through social media, as we developed infrastructure and appropriate methods to address basic research questions. At the individual level, we examined how personality, age, and gender were reflected on Facebook (Kern et al., 2014a; Kern et al., 2014b; Schwartz et al., 2013b). At the U.S. county level, we examined associations between language expressed on Twitter and life satisfaction and heart disease (Eichstaedt et al., 2015; Schwartz et al., 2013a).

Our Facebook data were drawn from the MyPersonality dataset (Kosinski, Stillwell, & Graepel, 2013). MyPersonality was a Facebook application that allowed users to take a personality test, based on the International Personality Item Pool (IPIP; Goldberg et al., 2006), and assessed personality based on the Big Five (extraversion, agreeableness, conscientiousness, neuroticism, openness). Users completed between 20 and 100 items representing the five factors. Users could optionally share their Facebook status updates for research purposes; with permission, their entire Facebook feed was downloaded and linked to their personality scores, and then identifiers were automatically removed. The resulting database of 20 million status updates was selectively made available to researchers for secondary data analysis. Here we use data from about 70,000 users to illustrate processes involved in working with social media language data.

Selecting a Dataset

Before any analyses can be performed, data must be obtained. Careful consideration should be given toward which data will be most appropriate for the question at hand, and whether informative data are available and accessible. As Borgman (2015) noted, "having the right data is usually better than having more data" (p. 4). Some shared data resources exist. For instance, the MyPersonality application includes Facebook status updates, personality scores, and some demographic information. Many social media platforms allow researchers to access data, although costs and the amount of accessible information vary. Volunteers can be recruited to share their private data (e.g., via Facebook), but such an approach can be difficult and expensive. Other data are simply not accessible; companies such as Google and Microsoft will not share private information such as search query data or emails. As public concerns about privacy continue to evolve, regulations and possibilities for data access will continue to ebb and flow. When planning a study, it may be helpful to be flexible in terms of the platform used and questions asked.

A certain amount of data per unit of observation is needed, especially when developing a language model. Language is noisy, and analyses are made harder by ambiguities, multiple word senses and uses, and slang. Similar to the need for multiple items on a self-report measure, a minimal number of words are needed to reduce noise from sparse responses. There tends to be considerable variation between users, and with a small number of users, models will over-fit to the sample, reducing generalizability. A single post will rarely have sufficient words to build a stable model. To address this, we pool language across available "documents" (e.g., Facebook status messages or tweets) to create a broader sampling of a person's or a group's language, combining all language shared by a user. Once a language model has been built on the basis of more words from many users, it can be applied to the language of users who have fewer words. Still, as responses become sparse, the accuracy of the model diminishes considerably.

In general, it is better to have more words per person and a greater number of persons. We generally use 1,000 words as a minimal criterion. To test this criterion, we calculated the effect of word count on accuracy of our models. **Figure 1** shows how error varies according to the number of words available for age and extraversion, across 4,000 randomly selected MyPersonality users. The x-axis is total words written (logarithmically scaled), the y-axis is the mean absolute error, the line on each graph was fit with LOESS regression (Cleveland, 1979) and the shaded area indicates the 95% confidence interval. For both age and extraversion, the graphs remain relatively flat after 1,000 words, although for extraversion, 500 words may be sufficient. The graphs suggest that is preferable to have more users to build a model over, rather than having fewer users with more language.

A related issue is the availability of outcome data. Analyses on different levels (e.g., individuals, group, regions, etc.) require different sorts of data. To examine Big Five personality and word use, we had 70,000 individuals with at least 1,000 words available. For communities, we find U.S. counties with at least 50,000 words available to be a good unit of analysis, as this provides several thousand units of analysis (and many degrees of freedom), compared to U.S. states (with fewer than 50 degrees of freedom), which tend to give many spurious results. Fortunately, there is a growing trend to make datasets more commonly available, particularly

for county or region level data, where there are fewer privacy concerns than with individual data (e.g., www.data.gov, www.data-archive.ac.uk, http://www.countyhealthrankings.org/, www.icpsr.umich.edu, http://www.cdc.gov/DataStatistics/, www.mypersonality.org/wiki). Analyses will benefit from different datasets being connected together, although such integration raises ethical issues.

Extracting Data

Once a dataset is selected, the data need to be downloaded. Social media data are generally accessed through an *Application Programming Interface* (API), which is a format that specifies and structures data, and provides an associated syntax that allows computer programs to communicate with one another. APIs are like translators, which allow application developers to create applications on their own systems and then seamlessly share the content with users, enhancing the user's social media experience. APIs also allow analysts to pull information without disrupting users' experiences. APIs also make it easy to handle large amounts of streaming (live) data in ways that would not be convenient through a web browser or other download method.

For example, Twitter makes a random sample of all public tweets available in real time, which can be accessed through the Twitter API (https://dev.twitter.com/streaming/public). To access this, you need to have a Twitter account. Upon registering, you receive an API key, API secret, access token, and access secret. You next create a blank Twitter application, which is used to retrieve the data, and then indicate what data you would like. This can include tweets, user information, entities (meta-data and contextual information), and places. As part of the code, a destination for the data is specified, such as a CSV file or database.[1] Other social media such as Weibo (the Chinese analog to Twitter) provide similar APIs.

Twitter includes limits on how much information you can request each hour as a free user (1% random feed per day; alternatively, one can retrieve data based on a specific criterion, such as geographic location). As the size of the data increases (to the order of four billion tweets/ day), this can quickly overwhelm a single computer, so hardware demands require planning and monitoring.

One often wants "meta-data" about each social media post, such as the time it was posted, location, who posted it, and the user's age, gender, and ethnicity. Some information (e.g., time of posting) is easy to extract through the API; other information can be inferred from user profiles. For instance, only a small percentage of tweets come with latitude/longitude coordinates. Twitter user profiles include an optional free-response location field, which we have used to infer location.

To illustrate, we mapped tweets to U.S. counties (Schwartz et al., 2013a). Twitter data were drawn from a random set of users collected from June 2009 to March 2010. Of the collected tweets, 148 million tweets could be mapped to U.S. counties. To map the tweets, coordinates were easily mapped. If the city and state were noted, the county could be determined. When only the city was included, we excluded large non-U.S. cities such as London

[1] See for instance https://spring.io/guides/gs/accessing-twitter/ or http://mike.teczno.com/notes/streaming-data-from-twitter.html for detailed instructions on accessing tweets.

or Paris, and ambiguous U.S. cities. Phoenix is most likely Phoenix, Arizona, whereas Portland could be Portland, Oregon or Portland, Maine, and was thus excluded. Such an approach produced fewer incorrect mappings at the cost of being able to map fewer tweets. Human raters checked a subset of the mappings for accuracy; 93% of those mapped were judged to be correct.

We typically use these location estimates to make geographically specific predictions, which are ultimately validated against a more reliable geographic dataset (e.g., U.S. Census data). To further quantify the error of using the free-response location as an estimate, we predicted county-level life satisfaction (based on self-reported data) and compared the accuracy of the model (*r*) based on location field reports to the proportion of geo-coded tweets. This quantifies how much using text versus geolocation affects things, but ignores any systematic differences between geocodes (i.e., latitude & longitude) and text-geolocated tweets.[2] As illustrated in **Figure 2**, the average error was *r* = .05 in the uncontrolled model and *r* = .04 in a model controlling for demographics. This suggests that the 7% inaccuracy does not appear to be causing systematic differences in prediction performance.

Some researchers only use data where the geolocation can be confirmed (e.g., Cheng & Wicks, 2014; Helwig, Gao, Wang, & Ma, 2015; Lampos & Cristianini, 2010). However, many more tweets that can be mapped to counties from the free-response location field than from geocoded coordinates (15 to 25% versus 2 to 3%), which allows more fine-grained analyses in space and time. One does not need geocoded coordinates to validate these; one only needs to establish the error rate over a random sample. Numerous approaches for inferring geolocation in social media have been used, ranging from simply keeping the roughly 2% that have precise locations on them, to noting that people tend to be close to the median location of their friends and followers (e.g., Backstrom, Sun, & Marlow, 2010; Bo, Cook, & Baldwin, 2012; Cheng, Caverlee, & Lee, 2010; Jurgens, 2013; Kinsella, Murdock, & O'Hare, 2011; Kong, Liu, & Huang, 2014; Li, Wang, & Chang, 2012; Mahmud, Nichols, & Drews, 2012; McGee, Caverlee, & Cheng, 2013; Rout, Preotiuc-Pietro, Bontcheva, & Cohn, 2013). A systematic review and comparison across nine methods found considerable differences in performance, and suggested that although self-reported location has been useful, it is less accurate in recent Twitter data (Jurgens et al., 2015). As social media is a dynamic system, best practices for geolocation will remain an active area of research.

Preparing Data for Analysis

After obtaining a dataset, information needs to be extracted and converted into a usable form. With big data, this stage can take considerable time, programming skills, and computing resources. It is particularly helpful to work with a computer scientist at this stage.

[2] Specifically, for counties that tweeted at least 50,000 words (*N* = 1071, 148 million tweets), we trained a ridge regression model to predict life satisfaction using 1-, 2-, and 3-gram features plus 2000 topic features applied on the county level. Parameters of the regression model were tuned using 10-fold cross validation, with 90% of counties used for training and 10% for testing. Each county was in the test group once, producing an out-of-sample model-predicted life satisfaction score for each county. We repeated this including demographic measures as additional features. We then calculated the error between the predicted scores and the survey-measured life satisfaction scores (with and without demographics).

With smaller language data, it is possible to directly create analytic features using a closed vocabulary tool. However, care should be taken to capture the oddities of social media expression.

Tokenization. The data accessed through the API form a database with language data (social media posts and their metadata) and associated outcome variables, either at the individual, group, or region level. The language data in its raw form is ill suited for quantitative analysis – it is just a sequence of characters. *Tokenization* refers to the process of splitting posts or sentences into meaningful tokens or words, which may be known dictionary words, misspellings, punctuation, netspeak (e.g., lol, brb), emoticons (e.g., "<3" is a heart, ":)" is a smiling face), and other variations. Sequences of letters are automatically identified, with adjustments made to separate punctuation from words. This is trickier than it seems, as "yesterday, I" is three tokens (the comma is not part of the word "yesterday", while "1,200" is one token, as is the emoticon ";-)").

The tokenizer needs to be sensitive to the misuse of language common in social media. For example, "dis sux... wonder who i can share dis with... dis kinda misery needs company" includes multiple misspellings, slang, and ellipses. A good tokenizer will break this into "dis" "sux" "…" wonder" "who" "I" "can" "share" "dis" "with" "..." "dis" "kinda" "misery" "needs" "company". Fortunately, good tokenizers are available (see sentiment.christopherpotts.net/code-data/happyfuntokenizing.py, our improvement on it: http://www.wwbp.org/data.html, or http://www.ark.cs.cmu.edu/TweetNLP/, and http://nlp.stanford.edu/software/tokenizer.shtml).

Analysts often combine apparently similar tokens, treating for instance "don't" and "dont" or "like" and "likes" as equivalent. Similarly, one can automatically normalize (i.e., translate into standard English) both the words (Han & Baldwin, 2011) and syntax (Kaufmann & Kalita, 2010). However, such combinations and translations should be done with caution, as such differences can reveal individual characteristics. For example, the use of the apostrophes in contractions correlates with (high) neuroticism and (low) openness to experience. Similarly, use of "sleepin" rather than "sleeping" reveals socioeconomic status. It is unclear how to translate emoticons, and translations rarely capture the true spirit of the original. Translating "girrrls" to "girls" or "boyz" to boys" may keep the meaning, but loses the connotations and emotional content. It is also common to remove "stop words" – words like "the" or "a". This can be useful in smaller datasets, especially when the focus is on classifying patterns within the text (e.g., most common words across a corpora). However, for individual differences, removing stop words is often ill-advised, as use of determiners correlates with both age and personality (Pennebaker, 2011; Schwartz et al., 2013b). It is often preferable to simply process the non-normalized tweets, counting or parsing the "words" and emoticons (Kong et al., 2014).

Stemming. One possibility for data preparation is *stemming*, in which words sharing a common stem are mapped to that stem (Porter, 1980). For instance, "sleep, sleeps and sleeping" would all be replaced by "sleep." This is generally not advisable with large datasets, as it tends to lose word distinctions that are often informative; "sleeping" is not exactly the same as "sleep", and different uses of the same stem might reflect important user characteristics. Tools like LIWC that use pattern matching are even worse, for example

collapsing “treasure” and “treasury” into a pattern of words that begin with “treasur*”. However, such simplifications may be useful for small datasets.

Multi-word expressions. Individual words suffer from ambiguity (e.g., is “tender” a feeling, a characteristic of steak, or a financial term?); their meaning depends strongly on the context. There is a large field of word sense disambiguation that attempts to address these problems (Navigli, 2009), but an easier, and highly effective solution, is to collect multi-word expressions (Sag, Baldwin, Bond, Copestake, & Flickinger, 2002). Short sequences of words that commonly occur together (e.g., “happy birthday”, “4th of July”) can be automatically identified, allowing for more context-sensitive analyses (Finlayson & Kulkarni, 2011). We tend to only use 2-grams (two adjacent words, or *bigrams*) and 3-grams (three adjacent words, or *trigrams*); longer phrases offer little benefit, as their individual occurrence rates are very low.

We identify and select informative 2-grams and 3-grams using the pointwise mutual information (PMI; Church & Hanks, 1990; Lin, 1998):

$$pmi(phrase) = log\frac{p(phrase)}{\prod_{word \in phrase} p(word)} \quad \text{Eq. 1}$$

The PMI is the logarithm of the ratio of the observed probability of two or three words co-occurring together, *p*(phrase), to what the probability of the phrase would be if the probabilities of the words in it were statistically independent (i.e., the product of their independent probabilities). The word probabilities, *p*(word), are simply the count of each word (*count*(word)) or phrase (*count*(phrase)) divided by the total number of words in the dataset (*N_words*):[3]

$$p(word) = \frac{count(word)}{N\_words} \quad \text{Eq. 2a}$$

$$p(phrase) = \frac{count(phrase)}{N\_words} \quad \text{Eq. 2b}$$

PMI bigrams help reduce word sense ambiguity – “sick of” is not the same as “sick”, just as “hot dog” is not a kind of “dog.” Positive or negative PMIs indicate that the words co-occur more or less often (respectively) than would occur by chance, and are more useful than simply picking pairs of words that frequently occur together. For example, the sequence of words “New”, “York”, and “City” will occur much more often than one would expect if they were independent; thus, *p*(phrase), the numerator in PMI, will be much larger than the product of all three individual word probabilities, the denominator in PMI, and a large positive value will result.

[3] Technically, *p*(word) and *p*(phrase) are maximum likelihood estimates of the *p* parameter of a Bernoulli distribution. In theory, if count(phrase) = 0, then PMI would not be defined. In practice, one never applies PMI to a phrase that does not occur.

Researchers have the option to observe phrases at various PMI thresholds and tune this parameter to their liking. We typically keep the two- and three-word phrases that have a PMI value greater than 1.5 times the number of words in the phrase (i.e., with 2-grams, we select phrases with a PMI greater than 3). Higher values limit phrases to only those that are most meaningful, while lower thresholds allow one to capture nuanced sequences, which may be helpful for prediction.

Labeling. In preparing the data for further analysis, it is often necessary to annotate or label the social media messages. Amazon's Mechanical Turk (Mturk, https://www.mturk.com/mturk/welcome) currently is a good platform for such tasks, as workers can be paid a minimal amount to label messages (Buhrmester, Kwang, & Gosling, 2011; Mohammad & Turney, 2013). For researchers outside of the U.S., Prolific Academic (https://prolificacademic.co.uk) provides an alternative.

For instance, messages might be labeled for signals that will change the resulting analyses. **Table 1** summarizes various discrepancy labels, with examples as to how they might apply to different types of text. Messages can also be labeled for the extent to which they indicate a particular construct. For example, we had messages rated for several aspects of well-being (positive emotion, engagement, relationships, meaning, accomplishment). After reading brief definitions, raters were randomly shown Twitter or Facebook messages and asked to indicate the extent to which the message indicated each category. Raters read the whole message, such that words could be considered in the full context that they occurred.

As with any other rating process, several raters should label messages to ensure adequate reliability. We typically use three raters and calculate the intraclass correlation coefficient (Shrout & Fleiss, 1979) as a measure of rater agreement, with the average of the three ratings as the final message score. For example, three raters annotated 6000 messages for their temporal orientation (i.e., if language is past, present, or future oriented), which took about 150 human hours, with an inter-rater reliability of .85 (Schwartz et al., 2015).

Grouping Words: Closed and Open Vocabulary Approaches

Whether the purpose is to describe patterns in the data or to make predictions, tokens need to be converted into numbers, such as the frequency that certain words or categories occur. Various approaches have been developed to group similar words together. Psychological studies of language have typically used *closed-vocabulary approaches*, in which data are passed through a pre-defined lexicon (or *dictionary*; i.e., a list of related words), which are developed *a priori*. Methods from computer science enable *open-vocabulary approaches*, which allow topics or groups of words and symbols to emerge from the data. Open-vocabulary approaches are not limited to preconceived notions of a particular topic and can accommodate unconventional language that is quite typical of social media data. Such methods can substantially improve predictions of various outcomes. However, sufficient data are needed, and the results can be harder to interpret.

In practice, closed vocabulary approaches are easiest for psychologists to implement and are often more practical. For a psychologist with several hundred individuals who have shared their social media data and completed questionnaires, closed-vocabulary approaches

can derive scores for established content categories, but there are insufficient data points for open-vocabulary approaches. With more data, a combination of closed and open approaches can be used, providing multiple approaches for honing in on consistent patterns. For instance, Yarkoni (2010) examined the personalities of bloggers, examining word categories that correlate with the Big Five factors (a closed vocabulary approach) and words correlating with each factor (an open vocabulary approach). There are a growing number of methods that allow a combination of open and closed vocabulary approaches, such as zLable Latent Dirichlet Allocation (LDA), supervised LDA, word embeddings, and skip gram modeling (e.g., Andrzejewski & Zhu, 2009; Bengio, Ducharme, Vincent, & Jauvin, 2003; Blei & McAuliffe, 2007; Collobert & Weston, 2008; Mikolov, Chen, Corrado, & Dean, 2013; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Turian, Ratinov, & Bengio, 2010).

Closed vocabulary approaches. Closed-vocabulary approaches are widely used in social media analysis. By applying *a priori* created lexica across thousands of Facebook users and blogs and millions of word instances, extraversion related to using more positive emotion words, whereas neuroticism related to using more negative emotion and swear words (Gill, Nowson, & Oberlander, 2009; Sumner, Byers, & Shearing, 2011). In over 140 million words from nearly 20,000 blogs, older and male bloggers tended to use more words related to religion, politics, business, and the Internet, whereas younger and female bloggers used more personal pronouns, conjunctions, fun, romance, and swear words (Argamon, Koppel, Pennebaker, & Schler, 2007). Across 16,000 Twitter users and two million tweets, Christians used more religious, positive emotion, and social process words, whereas atheists used more negative emotion and insight words (Ritter, Preston, & Hernandez, 2014). In millions of Facebook posts, positive and negative emotion expressions related to local weather reports (Coviello et al., 2014).

In psychological research, closed-vocabulary approaches have most commonly been implemented through the Linguistic Inquiry and Word Count program (LIWC; Pennebaker, Chung, Ireland, Gonzales, & Booth, 2007). LIWC was developed to capture multiple psychological dimensions (Pennebaker & King, 1999). Words were compiled from dictionaries, thesauri, existing questionnaires, and other sources, and then three or more judges independently rated whether the words should be included in each category (Pennebaker & Francis, 1996; 1999). The current version includes 64 different categories, within nine main types (affective, social, sensory, biological, and cognitive processes; verbs; relativity, function and miscellaneous words), and has been translated into 12 languages (see www.liwc.net for more information).

Though a relatively simple interface, LIWC allows text to be turned into numeric values. The program passes text through a “processor” (i.e., tokenizer and word counter), and provides the frequency that a user mentions each category. Frequencies should be adjusted by the total number of words, as users differ in the number of words that they write, (which is also the probability that any random word in a document belongs to the given category):

$$p(category) = \sum_{word \in category} p(word) = \frac{\sum_{word \in category} count(word)}{N\_words}$$

Eq. 3

The relative frequencies can then be summarized descriptively, correlated with other variables, or used as predictors in a regression analysis.

To illustrate, using the MyPersonality dataset, **Table 2** summarizes descriptives for the social, affective, and cognitive LIWC categories, and significant (Bonferroni-corrected) correlations with extraversion. Analyses were done using our custom Python codebase, part of which we have released open-source (see wwbp.org/data.html). The LIWC dictionaries are included in the LIWC software (available with cost from liwc.net), which we loaded into a suitable mySQL database table. After tokenization, we matched the tokens in the status updates (also stored in mySQL) against the words included in the LIWC dictionaries, and calculated the overall relative frequency of LIWC dictionaries for all users, which yielded the descriptives reported here. These relative frequencies were then regressed on self-reported extraversion scores (yielding β_e), with age and gender included as controls (β's not reported here). Positive emotion was most strongly positively related to extraversion, and insight, certainty, and negative emotion words were inversely correlated.

Researchers can also develop their own lexica. A psychologist might begin with a theory, create a list of words representing that theory, use judges to rate words for relevance, and then validate the lexicon against relevant criteria. Such a list can then be added to existing lexica programs (e.g., LIWC), taking advantage of the same word count infrastructure. For example, Cohen (2012) developed and provided preliminary evidence for a lexicon for cognitive rigidity. A balance must be found between capturing alternative spellings and word endings, while not capturing too many irrelevant words. This is another area where the integration of psychology and computer science is useful, as lexica can be expanded and improved using a *supervised learning* approach (e.g., Fernando, Fromon, Muselet, & Sebban, 2011; Lian, Wang, Lu, & Zhang, 2010; Mairal, Bach, Ponce, Sapiro, & Zisserman, 2009; Schütze & Pedersen, 1997). For example, words that co-occur might be automatically identified. Human raters indicate the extent to which the automatically extracted "similar" words are used in accordance with the lexicon definition. Incorrect uses are then fed back into the algorithms, improving lexicon quality.

Open vocabulary approaches. Although closed-vocabulary approaches provide psychologists with tools for analyzing social media and other language data in a structured manner, the real power comes from bottom-up approaches that allow the data to tell their own stories. Closed vocabulary approaches typically rely on dozens to hundreds of single words. Statistical and machine learning methods can process tens of thousands of words and phrases to find those most correlated with a trait, behavior, or outcome of interest.

Latent Semantic Analysis (LSA). The original approach to extracting meaningful information from large semantic datasets was *latent semantic analysis* (LSA; Deerwester et al., 1990). LSA is a method of dimensional reduction applied to the matrix created by documents (e.g., Facebook statuses) as rows and words as columns, with entries capturing the relative

frequency of occurrence of a word in a given document. LSA applies Singular Value Decomposition to this matrix, such that documents can be represented as a distribution of latent semantic factors (akin to factor analysis). The distance of different documents in the latent semantic space is derived by calculating the cosine similarity of the vectors of the two documents giving their loadings on the factors (Wolfe & Goldman, 2003; for a fuller discussion, see Landauer & Dumais, 1997).

Due to how word similarity is captured, LSA and other generally distributed representation methods such as word2vec (Mikolov et al., 2013) and GloVe (Pennington, Socher, & Manning, 2014) are good for measuring document or word similarity, and have been used to look at deeper relations between words and documents. LSA is a particularly suitable method for generating an automated distance metric between documents, which for example could be used to automatically score test responses by considering the distance between a student's answer and the correct reference answer. However, for psychological characteristics such as personality, the resulting factors themselves are not easily interpretable, as words that load highly on a given factor are not necessarily semantically coherent. Rather, it is helpful to find sets of words (i.e., "topics" or "clusters"). One could cluster the word2vec representations or use LDA topics that capture semantically related words, the latter which we focus on here.[4]

Differential Language Analysis (DLA). In our own work, we have used a *differential language analysis* (DLA; Schwartz et al., 2013b) approach. DLA finds words and phrases that most strongly correlate with a given characteristic. **Figure 3** illustrates the DLA process. Stemming from an ordinary least squares regression framework, thousands of regression analyses are run, one for each language feature (e.g., word, multiword phrase, topic) independently. In each of these regressions, the relative frequency of a language feature is the predictor and the characteristic of interest (e.g., extraversion) is the outcome variable:

$$\hat{y} = \beta_0 + relfreq_{word}\beta_1 + cv_1\beta_2 + cv_2\beta_3 + cv_k\beta_{k+1} + \epsilon \quad \text{Eq. 4}$$

where $\hat{y}$ is the outcome of interest, $relfreq_{word}$ is the relative frequency of a language feature, and $\{cv_1, \ldots, cv_k\}$ are any number of control variables. All variables are standardized (mean centered and divided by the standard deviation), and β_0 is often 0. We generally control at least for age and gender. The resulting standardized parameter estimates, which we report as partial correlation coefficients (β_i), indicate the strength and direction of the unique association between the word/phrase with the outcome, holding the other covariates constant.

Typical datasets contain tens of thousands of words and phrases. As differing amounts of text are usually available for each person (or county), we adjust the frequencies for each user by the total number of words and phrases that a person used, deriving the relative frequency of that word. The resulting frequency distributions tend to be extremely positively skewed, with many zero values and a few outlying individuals. We first remove n-grams that are not used by at least 1% of the sample. Then, to reduce the impact of outliers, we transform the n-gram distributions using the Anscombe (1948) transformation:

[4] For a full discussion and comparison to LSA, see Griffiths, Steyvers, & Tenenbaum, 2007. For an excellent worked example of the application of LDA to couple's therapy transcripts, see Atkins et al., 2012.

$$2\sqrt{p(phrase) + \frac{3}{8}} \qquad \text{Eq. 5}$$

This results in the adjusted relative frequency for each word or phrase with a more stable variance.

This analysis results in thousands of word/ outcome correlations, most of which are very small in magnitude (ranging from r = .00 to .20). As a heuristic for separating signal from noise, we calculate Bonferonni-corrected p values, and only consider the estimates as potential signal when the corrected p value is less than .05 (or .001 for a stricter criterion). For example, with 20,000 language features, we retained p values less than 0.001 / 20,000, or $p < .00000005$ (Schwartz et al., 2013b). This is the most conservative form of correction; less conservative approaches like the Benjamini-Hochberg can also be used (Benjamini & Hochberg, 1995). Alternatively, the split-half reliability between two sets of data can test the robustness of effects. Many associations may still reflect chance. It is important to read through the results and see if they make sense, and to be wary of over-interpretation of single significant words. In addition, cross-validation is key to not over-fitting the model (see below).

As a final step, we use a modified word cloud, created using the advanced version of Wordle (www.wordle.net/advanced), to visualize the resulting correlations. We use the *size* of the word to indicate the strength of the correlation, and *color* to indicate the frequency that the word is used. This results in a single image with two dimensions of information (frequency and correlation strength) that illustrates the 50 to 100 words and phrases most strongly correlated with the outcome. We found that expert and lay audiences alike can quickly derive an intuitive sense of the results presented in this way. To illustrate, **Figure 4** visualizes the words and phrases most strongly positively and negatively correlated with extraversion. Individuals high in extraversion used words such as “party”, “chillin”, “love you”, and “can’t wait”. Interestingly, low extraversion (introversion) speaks to computer-oriented introverts, with words such as “computer” and “anime”. [5]

***Automatic topic creation*.** Various techniques make it possible to automatically generate categories or topics, based upon words that naturally cluster together, similar to latent class cluster analyses (Clogg, 1995). One common approach is Latent Dirichlet Allocation (LDA; Blei, Ng, & Jordan, 2003), which can be performed using the Mallet package (MacCallum, 2002).[6] Considering the entire distribution of messages (across users), an algorithm iterates through the words and finds those that commonly occur in the same posts. Words receive weights according to how much they load on each topic, just as items load on latent variables in

[5] At this point, www.lexhub.org/tools currently runs lexica and weighted lexica, but there is not an easy way for readers to run DLA. Online tools for running DLA will be available on this site in the future.

[6] Note that we describe one type of LDA here, but there is a huge range of LDA variations. Indeed, over 1,000 papers exist describing different ways of building prior knowledge into LDA, such as methods that automatically select the number of clusters, use of lists of words of different type, or make use of the fact that word selection is driven both by the topic and by who the author of each document is (e.g., Blei, 2012; Blei & Lafferty, 2007; Doyle & Elkan, 2009; Li & McCallum, 2006; Paul & Dredze, 2011; Rosen-Zvi, Griffiths, Steyvers, & Smyth, 2004; Teh, Jordan, Beal, & Blei, 2006; Wallach, 2006; Wang, Thiesson, Meek, & Blei, 2009; Zhu, Ahmed, & Xing, 2012).

factor analysis. Topics are non-orthogonal, such that words can occur in multiple topics, reflecting that words can have multiple senses.

As iterations could continue on endlessly, it is best to specify a set number of topics beforehand. We have found that there is a trade-off between precision and redundancy. As the number of topics increases, interpretation becomes easier, as the topics are more precise in their coverage, but so do the number of redundant topics – those that seemingly cover the same qualitative concepts. For example, in the MyPersonality dataset, we generated 50, 500, and 2000 topics. **Table 3** notes topics where the words "happy" and "play" were among the top 10 words of the topics. Happy appeared in two, eight, and 20 topics respectively. As the number of topics increases, holidays increasingly split across topics (e.g., a single holiday topic amongst the 50 topics, versus separate topics for Easter, Thanksgiving, Valentine's day, etc. amongst the 2000 topics).

More topics potentially could be extracted, but we stopped at 2000 to keep the number of topics manageable (full set of topics available from http://www.wwbp.org/data.html). The topics could also be further clustered together into higher-order topics. Facebook statuses and tweets are quite general and often written by a heterogeneous set of users; in situations where the domain of the text is more limited (e.g., prompted essays) or the sample comes from a more homogenous group (e.g., undergraduates at a single university), one may find that a smaller number of topics are sufficient.

The resulting topics do not automatically have labels. Human raters can read through the words and provide labels that seemingly best represent the topic, but the labels are somewhat arbitrary. For example, in one topic, the strongest word was "food", and other words included "Chinese", "restaurant", "Mexican", "dinner", and "eat". This could easily be labeled a food topic. In other cases, the words together suggest meaning that goes beyond any single word within the topics. For instance, a topic included the words: "money", "support", "donate", "donations", "raise", and "Haiti", pointing to a philanthropy topic, even though the word "philanthropy" was not directly used.

We then calculate the probability that a person uses each topic as:

$$p(topic) = \sum_{word \in topic} p(word)p(topic|word) \qquad \text{Eq. 6}$$

where $p(word)$ refers to the normalized word use by a given person, and $p(\text{topic}|\text{word})$ refers to the probability of a topic given the word, provided by the LDA procedure.

The resulting probabilities could be visualized or used as features in other analyses; just like the words and phrases, they express the language of a user as a distribution over topics. Applying the 2000 topics to the MyPersonality data, **Figure 5** visualizes topics that were most strongly positively and negatively associated with extraversion. We used the size of the word to indicate its weight within the topic, rather than the strength of the correlation coefficient. Thus, the larger the word, the more it represents that topic.

Tools for Analysis

Across the process of extracting and preparing data, many different tools and analytic packages are available. Programmers often use Python or another programming languages to

extract and process text. An extensive number of articles and books on text mining are available (e.g., Agarral & Zhai, 2012; Grossman & Frieder, 2012; Jurafsky & Martin, 2014).

A growing number of tools can be used to extract meaningful information. Beyond LIWC, SAS sentiment analysis (http://www.sas.com/en_us/software/analytics/sentiment-analysis.html) determines positive and negative sentiment in a set of text. General Inquirer (http://www.wjh.harvard.edu/~inquirer/Home.html), first developed by Harvard University in the 1960s, includes dictionaries centered on psychological and sociological theories, including deference, welfare, and decision-making. DICTION (http://www.dictionsoftware.com) was developed to analyze political texts, and includes five "master categories" (certainty, activity, optimism, realism, and communality), based on combinations of 35 dictionaries and sets of language statistics (e.g., fraction of words longer than five letters). Lexhub (http://lexhub.org) highlights packages and tools that might be helpful. For smaller language data, various programs have been developed to assist with qualitative data analysis (see https://en.wikipedia.org/wiki/Computer-assisted_qualitative_data_analysis_software for a listing of different options). The best lexica or analytic program to use depends on the type of data available and the research questions.

Once social media data are processed, the resulting output is typically in the form of a comma separated value (CSV) file, which can be used as a dataset in statistical analytic programs such as R, STATA, or SAS. Excel and SPSS have trouble opening and working with very big files, so tools like R and SKLearn are generally preferable.[7]

Obstacles and Challenges

Any number of analyses could be applied to the features generated through closed and/or open approaches to describe, visualize, and make predictions from the data. A detailed description of such methods is beyond the scope of this article. Regardless of the methods used, numerous challenges and issues arise through the analytic process, many of which are different from the problems and controversies encountered in traditional psychological studies. In this last section, we highlight key issues related to processing and analyzing data, interpreting results, and ethical considerations (see also Grimmer & Stewart, 2013; Iliev, Dehghani, & Sagi, 2014; Shah, Cappella, & Neuman, 2015; Yarkoni, 2012 for methodologies and discussions of issues).

Challenges with Processing and Analyzing Data

Memory and storage. For the initial processing of data, some sort of server or database management system (DBMS) is needed to store the raw data. Size needs to be considered. From 2012 to 2013, we extracted almost 2 billion tweets from Twitter's random sample feed. Storing the tweets and their tokenized forms required over one terabyte of storage space – not a problem for a modern desktop computer, but challenging for a laptop. Working memory can also be a problem. Doing queries of the form "give me all tweets that contain the word 'anger'

[7] See scikit-learn (http://scikit-learn.org/stable/) for open source implementations and documentation of several forms of regularized regression, and https://github.com/scikit-learn/scikit-learn for source code. R packages such as https://cran.r-project.org/web/packages/glmnet/ are also useful.

but not the expression 'anger management'" can be painfully slow without a reasonable database. Such issues can be addressed by creating a Hadoop cluster (http://hadoop.apache.org), which combines the computing power of multiple computers into a type of supercomputer. Each machine provides local storage of data, working memory and computing power, and the software combines it all together. This processing is increasingly done "in the cloud," for example through Amazon Web Services (AWS; http://aws.amazon.com/).

Language use and ambiguity. Language can be ambiguous, and consideration needs to be given to how to define "words". For example, in formal writing punctuation follows specific rules, whereas in social media it might reflect actual punctuation (e.g., a period ending a sentence), express emotion (i.e., emoticons), emphasize a point (e.g., …, !!!, !?!?), or reflect mistypes and misspellings. Although closed-vocabulary approaches make it relatively easy to count word occurrences, they ignore the context in which words are used and ambiguities that they point to. Open-vocabulary approaches can capture more of the context by identifying how the words are used, multiword expressions, and clustering similar words. Decision rules for tokenizing, parsing, and categorizing data need to be sensitive to multiple uses, and will need to evolve as people and platforms change.

Model error. A certain degree of error occurs and is carried throughout the process of preparing and analyzing data. When available, out-of-sample (cross-validated) predictions of measures derived through other methods should be used as a measure of external validity, and to give an upper bound to carried over errors. For instance, a model based on Twitter language of U.S. counties correlated with ground-truth population data collected by the Centers for Disease Control and Prevention at rates of $r = .42$, indicating that the model captured at least 17.6% of the variance in the heart disease rates (Eichstaedt et al., 2015). This does not tell us exactly where errors are introduced, but it does inform the uncertainty of the final estimator.

Due to relying on a Bayesian optimization technique, LDA results can be hard to reproduce (Lancichinetti et al., 2015). In practice, we estimate the topics once and then use those same topics in many different regression models. If one were to re-estimate the topics, one would need to re-run the regressions, as different topics would be found and hence different regression coefficients. However, the topic model itself is often not very theoretically interesting or important. Rather, we (and we think most researchers) are more interested in the repeated patterns of relationships between multiple topics and other outcomes (i.e., the predictive accuracy of the model). In line with the recent emphasis in the field (e.g., Duncan, Engel, Claessens, & Dowsett, 2014; Pashler & Wagenmakers, 2012), reproducibility is important, and we thus focus on broad patterns of results rather than any individual parameter estimate or statistical test. When we examined personality, we repeatedly found that topics with words related to positive emotion and social enthusiasm, expressed in various ways, were correlated with higher trait extraversion. This result does not rely on any single correlation or topic, and a similar pattern of results is found when just using single words or multi-word phrases in place of LDA topics, suggesting that the result is robust across different methods.

Over-fitting. One often wants to build a model that predicts an outcome based on a subset of predictors. However, the number of predictors is often far greater than the number of

observations available for analysis. Many of the predictors are highly correlated, redundant, and not used enough to be entered into the model, or used in such varied ways that the features add noise rather than signal. Standard ordinary least squares regression includes far too many features in the model. As a result, excellent model fit occurs by capitalizing on the many parameters at the expense of degrees of freedom, but such models do not generalize well to new data, and the coefficients are less likely to capture "real" associations between features and the outcome of interest.

Before any sort of regression models are run, the number of predictors should be pruned or reduced (Hastie, Tibshirani, & Friedman, 2009). Reductions are typically done in a training set (i.e., a random subset of the data), with final predictions done on a test set. This ensures that over-fitting is captured as a *poorer* fit on the test set, not just a *better* fit on the training set. A first approach involves removing features with minimal correlations to the target outcome, based upon the family-wise error rate (Guyon & Elisseeff, 2003). A second approach involves running a form of Principal Components Analysis (PCA; Hotelling, 1933; Martinsson, Rokhlin, & Tygert, 2011) separately for words, phrases, and topics, reducing the number of dimensions to ten percent of its original size. For example, in the MyPersonality data, we had over 51,000 features. By first removing features with minimal correlations and then running PCA, we reduced the number of features (i.e., predictors) to 5,100, a much better balance with the sample size (see Park et al., 2015).

Even with such reductions, machine learning methods may still converge on an overly specific solution that describes that training dataset but will not generalize to other data. One approach is to use *cross-validation*, a method for picking the values the give the best performance on a held out test set. A simple approach is to develop a model on one set of data (the *training* set), and then use the model to predict scores in a second independent set of data (the *test* set). A second approach involves a *k*-fold cross-validation (Hastie et al., 2009). Observations are randomly split into *k* similarly sized groups. One group is used as a test set, the others are used to develop the model. The groups are shuffled, and process is repeated until every group has been used as the test group once. The results are averaged together, providing an estimate of predictive accuracy.[8]

Regularization and variable selection. Despite reductions in the number of features, multicollinearity remains a problem. Further, some words are used equally often by most people and therefore have no signal. Other words and phrases are used extensively by a few users and rarely or not at all by the majority. This creates very positively skewed distributions with many zero values, violating assumptions of the OLS model. Methods are needed to stabilize the estimations.

The most common approach is *ridge regression* (or Tikhonov regularization; Hoerl & Kennard, 1970), which penalizes the sum of squares error and biases coefficients towards zero. Improved prediction is achieved, but bias increases, and as all predictors remain in the model, it is not a parsimonious model. An alternative is the Lasso method (least absolute shrinkage and

[8] Several implementations of cross-validation are freely available for evaluating a wide range of models, such as the well-documented, open source R package "caret" (Kuhn, 2015).

selection operator; Tibshirani, 1996), which penalizes the regression coefficients. As some parameters are driven to zero, they are removed from the model. However, it will only select as many variables (k) as there are number of cases (*N*). Also, if correlations among a group of variables are high, it will only select one variable from the group. A third alternative is elastic net regularization, which combines penalties from ridge and Lasso in a two-step process (Zou & Hastie, 2005). The process removes limitations on the number of variables, allows grouped effects, and stabilizes the L_1 regularization path.

As elastic net includes far fewer predictors in the model, it is generally the approach we recommend. Other methods are also possible, such as AIC and Mallow's Cp, but these tend to vastly overfit the data, putting in far too many features. One can also use a combination of L0 (a penalty proportional to the number of features selected) and L2 regularization, which often gives better models, especially when only a small fraction of the features will be selected, although at a higher computation cost.

Interpreting Results

The meaning of significant findings. There needs to be clear consideration of what estimated effect sizes actually mean. Unlike the typical psychological approach where a specific theory is tested, the computer iterates through the dataset and identifies correlations that may not simply be due to chance. We use a Bonferroni corrected *p* value as a heuristic for thinking about what associations may not be simply chance, but this does not mean that the identified words are anything other than noise. Even "strong" associations between words and characteristics tend to be small in size, according to conventional ways of thinking about effects. Language data can describe and make predictions about groups, but predictions derived from it tend to be quite noisy for any single individual.

Words versus topics. Similar to the way that multi-item scales tend to be more reliable than single-item measures (e.g., Diamantopoulos, Sarstedt, Fuchs, Wilczynski, & Kaiser, 2012), clusters or topics are often more informative than single words. Improved reliability increases expected effect sizes; whereas effect sizes for individual words tend to be small ($r < .10$), effect sizes for topics are often considerably larger (e.g., up to $r = .25$ for individual level factors and $r = .55$ for county level health factors). However, this is not always the case. For instance, we compared positive and negative emotion across six countries (U.S., United Kingdom, Canada, Australia, India, Singapore; Kern & Sap, 2015), first considering correlations with the LIWC positive and negative emotion categories, and then individual emotions. With categories, the dominant words were strikingly similar across the countries (e.g., "happy", "thanks", "please", "lol", "love", "good", and ":)"). There was greater distinction with single emotions, such as "anger" and "disgusting". Similarly, Grühn, Kotter-Grühn, and Röcke (2010) found that different trajectories characterized discrete emotions across the lifespan. The extent to which topics versus words should be used remains an area for future research, and psychological theory and interpretation will be key for distinguishing when each approach is most appropriate.

Fallacies. Ecological fallacies involve making conclusions about individuals based on grouped data (or more generally, making conclusions about phenomena at one level based on data from another level). Although we find that U.S. states with higher life satisfaction (e.g.,

Colorado) have higher rates of physical activity, this does not mean that a satisfied person is physically active. Even for sizable and significant correlations between 50 U.S. states and words, there are so many other explanatory factors that most interpretations are extremely weak at best, and just plain wrong at worst.

Exception fallacies can also occur, in which conclusions about groups are made based on exceptional cases. Certain users may use a single word vastly more than others. If models are not adjusted, models can be greatly influenced by outliers. It can even be the case that the most outlying cases are robots (i.e., automatic accounts set up to post information), such that conclusions could be based completely on non-humans.

We have found that one of the best guards against making these fallacies is to read through several hundred posts in which associations occur, to determine the extent to which conclusions make sense or are influenced by strange cases. For example, we examined words correlating with “pope”, as an attempt to measure religious affiliation. Surprisingly, correlated words included “scandal”, “Olivia”, “hood pope”, and “cassadee pope”. Reading through sample messages, it was clear that some cases of pope referred to Olivia Pope, a character in the television show “Scandal”, the song “It’s the Hood Pope” by artist A$AP FERG, or the singer Cassadee Pope, a popular American country music singer and songwriter. Irrelevant words and phrases could then be excluded, and the resulting word clouds clearly captured the Catholic pope that we initially intended, with words such as “Francis”, “Catholic”, “church”, and “Vatican”.

Non-representativeness of social media users. Although studies include a large number of people, those who post on Facebook or Twitter are a non-random sample. Further, people may post in a socially desirable fashion. Although true, these criticisms are less problematic than is often assumed. Most psychology studies employ non-random population samples (e.g., undergraduates), and many surveys suffer from desirability biases. Non-representative data is still valuable for understanding large populations (Shah et al., 2015), in the same way that survey research has been valuable for understanding various populations.

Facebook and Twitter users are not a representative sample of the population; older people are under-represented (but, interestingly enough, our volunteers seem to have a similar distribution of introverts and extraverts as the general population). Since we do have demographics of Facebook users, we can treat them as a stratified sample and adjust the data to reflect population statistics (Weeg et al., 2015). Still, the value of Twitter or other social media platforms as measures of community characteristics depends upon how much social media activity occurs in the community, with better signal coming from high-use areas (Farnham, Lahav, Monroy-Hernandez, & Spiro, 2015).

Desirability bias sounds like a worse problem, but most of our analyses speak to the fact that strong relative differences between individuals still occur. For example, even if introverts try to look a little more extraverted, on average they still post far less about parties and far more about reading. Finally, while self-censorship may occur, validating against alternative measurements still suggests we are capturing individual differences, and we still find enough posts with swear words or illegal drug use to warrant a “warning: offensive language to appear” in most of our presentations.

Ethical Considerations

A final important consideration is the ethics involved in any sort of social media study. A local university institutional review board (IRB) monitors all of our studies. Many corporations have ethics boards, but it is less clear who monitors the work that is done. Further, there is a growing need to determine the level of oversight that is appropriate for social media studies (Hayden, 2015).

With social media, it is almost impossible to completely de-identify people, and the information needs to be carefully secured from hackers. To keep data secure, we separate the client-facing server used by a Facebook application from the infrastructure that collects and protects identifiable user information (i.e., the secure data repository). The client-facing side is more at risk for code injection and other hacking attempts, so no identifiable information is stored within it. The secure server has the same access tokens (i.e., access credentials), but also pulls identifiable user information, in order to match information and connect available pieces of information. The secure repository is housed under the control of the University, which is strictly controlled at multiple levels.

Passing ethics review can seem like a frustrating process. Yet it catches potential harm that we might not see. For example, what information should be shown back to a user? We might think we are simply giving back to the user for giving us a bit of their time. But content can trigger underlying psychological issues. The researcher is removed from the participant and separated by numerous layers, making it challenging to determine if harm does occur. Further, some have suggested that Twitter or other social media platforms with publicly available data could be used to monitor characteristics such as well-being, illness, and political and economic trends. Yet it is questionable what users understand public to mean. Many users are unaware that their information might be used for research (Weinberg & Gordon, 2015). Researchers may need to protect people from what they do not know.

It is important to disclose to users in clear detail what data we are collecting and what we are using it for. In the case of experiments and manipulations, consent forms need to be very explicit and easy to understand, not buried within lengthy text. The ethical lines within both observational and experimental studies need to be constantly revisited as social media – and its users – evolve.

Conclusion

There is considerable value to studying behavior on social media platforms. Social media platforms represent different cultures, which are formed and reform over time. Social media language presents numerous opportunities, ranging from secondary analysis of existing information to real time monitoring of sentiment, health, and economic outcomes. Technology keeps evolving as computer scientists push the limits of what can be done. Psychologists play an important role in understanding the stories that arise from the data. As the novelty of big data wears off, a deeper layer of studies, which combine psychological theory with tools and methods from computer science will develop.

We have focused on textual analysis here, as we find that language is psychologically rich with information. Other mediums of communication can also be explored, such as sounds, pictures, and images. The power of observation comes when multiple sources and multiple

methods converge on similar claims. The amount of available data is inconceivable — people leave footprints everywhere of their moods, behaviors, personalities, and experiences. Social media has become a valuable part of social life, and there is much we can learn by collaboratively studying the tracks left behind, while being cautiously optimistic in our applications and approaches.

References

Aggarwal, C., C., & Zhai, C.X. (Eds.). (2012). *Mining text data*. New York: Springer.

Anderson, B., Fagan, P., Woodnutt, T., & Chamorro-Premuzic, T. (2012). Facebook psychology: Popular questions answered by research. *Psychology of Popular Media Culture, 1,* 23-37. http://dx.doi.org/10.1037/a0026452

Andrzejewski, D., & Zhu, X. (2009). Latent Dirichlet Allocation with topic-in-set knowledge. In *SemiSupLearn '09 Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing* (pp. 43-48). http://pages.cs.wisc.edu/~jerryzhu/pub/zlabel.pdf

Anscombe, F. J. (1948). The transformation of poisson, binomial and negative-binomial data. *Biometrika*, *35*, 246-254.

Argamon, S., Koppel, M., Pennebaker, J., & Schler, J. (2007). Mining the Blogosphere: Age, gender and the varieties of self-expression. *First Monday, 12*. http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2003/1878

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

Backstrom, L., Sun, E., & Marlow, C. (2010). Find me if you can: Improving geographical prediction with social and spatial proximity. In *Proceedings of WWW 2010* (pp. 61–70). http://cameronmarlow.com/media/backstrom-geographical-prediction_0.pdf

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), 57*, 289-300. http://www.jstor.org/stable/2346101

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin. C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research, 3*, 1137-1155. http://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf

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

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *Annals of Applied Statistics, 1*, 17-35. http://projecteuclid.org/euclid.aoas/1183143727

Blei, D. M. & McAuliffe, J. D. (2007). Supervised topic models. In *Advances in Neural Information Processing Systems 20 (NIPS 2007).* https://www.cs.princeton.edu/~blei/papers/BleiMcAuliffe2007.pdf

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent direichlet allocation. *Journal of Machine Learning Research, 3,* 993–1022. http://jmlr.org/papers/volume3/blei03a/blei03a.pdf

Block, J. (1993). Studying personality the long way. In D. C. Funder, R. D. Parke, C. Tomlinson-Keasey, & K. Widaman (Eds.), *Studying lives through time: Personality and development.* Washington, DC: American Psychological Association.

Bo, H., Cook, P., & Baldwin, T. (2012). Geolocation predication in social media data by finding location indicative words. *Proceedings of COLING 2012: Technical Papers* (pp. 1045-1062), Mumbai. http://www.aclweb.org/anthology/C12-1064

Booth, T., Mottus, R., Corley, J., Gow, A. J., Henderson, R. D., Maniega, S. M., …, & Deary, I. J. (in press). Personality, health, and brain integrity: The Lothian birth cohort study 1936. *Health Psychology*. http://dx.doi.org/10.1037/hea0000012

Borgman, C. L. (2015). *Big data, little data, no data: Scholarship in the networked world*. MIT Press.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's mechanical turk: A new source of inexpensive, yet high-quality data? *Perspectives on Psychological Science, 6*, 3-5. http://dx.doi.org/10.1177/1745691610393980

Chaix, B., Linstrom, M., Rosvall, M., & Merlo, J. (2008). Neighborhood social interactions and risk of acute myocardial infarction. *Journal of Epidemiology & Community Health, 62*, 62-68. http://dx.doi.org/10.1136/jech.2006.056960

Cheng, Z., Caverlee, J., & Lee, K. (2010). You are where you tweet: A content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management, CIKM '10* (pp. 759-768). Toronto, ON, Canada: ACM. http://faculty.cs.tamu.edu/caverlee/pubs/cheng10cikm.pdf

Cheng, T., & Wicks, T. (2014). Event detection using Twitter: A spatio-temporal approach. *PLoS One, 9*, e97807. http://dx.doi.org/10.1371/journal.pone.0097807

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computer Linguistics*, *16*, 22-29. http://www.aclweb.org/anthology/P89-1010.pdf

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association, 74,* 829-836. http://dx.doi.org/10.1080/01621459.1979.10481038

Clogg, C. C. (1995). Latent class models. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 311–359). New York, NY: Plenum Press. http://dx.doi.org/10.1007/978-1-4899-1292-3_6

Cohen, S. J. (2012). Construction and preliminary validation of a dictionary for cognitive rigidity: Linguistic markers of overconfidence and overgeneralization and their concomitant psychological distress. *Journal of Psycholinguistic Research, 41*, 347-370. http://dx.doi.org/10.1007/s10936-011-9196-9

Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *International Conference on Machine Learning, 2008.* http://ronan.collobert.com/pub/matos/2008_nlp_icml.pdf

Coviello, L., Sohn, Y., Kramer, A. D. I., Marlow, C., Franceschetti, M., Christakis, N. A., & Fowler, J. H. (2014). Detecting emotional contagion in massive social networks. *PLoS One, 9*(3), e90315. http://dx.doi.org/10.1371/journal.pone.0090315

Cox, M., & Ellsworth, D. (1997). Application-controlled demand paging for out-of-core visualization. *Proceedings of the 8th IEEE Visualization '97 Conference.* https://www.nas.nasa.gov/assets/pdf/techreports/1997/nas-97-010.pdf

Deerwester, S.C., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*, 391-407. http://www.cob.unt.edu/itds/faculty/evangelopoulos/dsci5910/LSA_Deerwester1990.pdf

Diamantopoulos, A., Sarstedt, M., Fuchs, C., Wilczynski, P., & Kaiser, S. (2012). Guidelines for choosing between multi-item and single item scales for construct measurement: A

predictive validity perspective. *Journal of the Academy of Marketing Science, 40*, 434-449. http://dx.doi.org/10.1007/s11747-011-0300-3

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

Doyle, G., & Elkan, C. (2009). Accounting for burstiness in topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning, 2009*. Montreal, Canada. http://cseweb.ucsd.edu/~elkan/TopicBurstiness.pdf

Duncan, G. J., Engel, M., Claessens, A., & Dowsett, C. J. (2014). Replication and robustness in developmental research. *Developmental Psychology, 50*, 2417-2425. http://dx.doi.org/10.1037/a0037996

Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., … & Seligman, M. E. (2015). Psychological language on Twitter predicts county-level heart disease mortality. *Psychological Science, 26,* 159-169 http://dx.doi.org/10.1177/0956797614557867

Farnham, S. D., Lahav, M., Monroy-Hernandez, A., & Spiro, E. (2015). *Neighborhood community well-being and social media*. Unpublished manuscript. Retrieved from http://thirdplacetechnologies.com/wp-content/uploads/2015/02/neighborhoodstudy_2_6_4.pdf

Fernando, B., Fromont, E., Muselet, D., & Sebban, M. (2011). Supervised learning of Gaussian mixture models for visual vocabulary generation. *Pattern Recognition, 45*, 897-907. http://dx.doi.org/10.1016/j.patcog.2011.07.021

Finlayson, M. A., & Kulkarni, N. (2011). Detecting multi-word expressions improves word sense disambiguation. *Proceeding MWE '11 Proceedings of the Workshop on Multiword Expressions: From Parsing and Generation to the Real World* (pp. 20-24). http://users.cis.fiu.edu/~markaf/doc/finlayson.2011.procmwe.8.20.pdf

Friedman, H. S. & Martin, L. R. (2011). *The longevity project: Surprising discoveries for health and long life from the landmark eight-decade study*. NY: Hudson Street Press.

Gill, A. (2004). *Personality and language: The projection and perception of personality in computer-mediated communication*. Doctoral dissertation, University of Edinburgh, United Kingdom. Retrieved from http://homepages.inf.ed.ac.uk/agill1/papers/GillAJ2003ThesisFinal.pdf

Gill, A.J., Nowson, S., & Oberlander, J. (2009). What are they blogging about? Personality, topic and motivation in Blogs. In *Proceedings of the Third International ICWSM Conference*. San Jose, CA.

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. http://dx.doi.org/10.1016/j.jrp.2005.08.007

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review, 114*, 211-244. http://dx.doi.org/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, 21*, 267-297. http://dx.doi.org/pan/mps028

Grossman, D. A., & Frieder, O. (2012). *Information retrieval: Algorithms and heuristics* (Vol. 15). Springer Science & Business Media.

Grühn, D., Kotter-Grühn, D., & Röcke C. (2010). Discrete affects across the adult lifespan: Evidence for multidimensionality and multidirectionality of affective experiences in young, middle-aged and older adults. *Journal of Research in Personality, 44*, 492-500. http://dx.doi.org/10.1016/j.jrp.2010.06.003

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3,* 1157-1182. http://www.jmlr.org/papers/volume3/guyon03a/guyon03a.pdf

Hampson, S. E., Edmonds, G. W., Goldberg, L. R., Dubanoski, J. P., & Hillier, T. A. (2013). Childhood conscientiousness relates to objectively measured adult physical health four decades later. *Health Psychology, 32,* 925-928. http://dx.doi.org/10.1037/a0031655

Han, B., & Baldwin, T. (2011). Lexical normalization of short text messages: Makn sens a #twitter. In *HLT '11 Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 368-378). http://www.aclweb.org/anthology/P11-1038

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

Hayden, E. C. (2015, April). Guidance is issued for US internet research [news blog]. *Nature, 496,* 411. http://dx.doi.org/10.1038/496411a

Helwig, N. E., Gao, Y., Wang, S., & Ma, P. (2015). Analyzing spatioemporal trends in social media data via smoothing spline analysis of variance. *Spatial Statistics, 14*, 491-504. http://dx.doi.org/10.1016/j.spasta.2015.09.002

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12,* 55–67. http://dx.doi.org/10.1080/00401706.1970.10488634

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology, 24*, 417-441. http://dx.doi.org/10.1037/h0071325

Jurafsky, D., & Martin, J. H. (2014). *Speech and language processing*. Pearson.

Jurgens, D. (2013). That's what friends are for: Inferring location in online social media platforms based on social relationships. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM).* https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/viewFile/6067/6366

Jurgens, D., Finnethy, T., McCorriston, J., Xu, Y. T., & Ruths, D. (2015). Geolocation prediction in Twitter using social networks: A critical analysis and review of current practice. In *Proceedings of the 9th International AAAI Conference on Web and Social Media (ICWSM).* http://cs.stanford.edu/~jurgens/docs/jurgens-et-al_icwsm-2015.pdf

Kaufmann, M., & Kalita, J. (2010). Syntactic normalization of twitter messages. In *Proceedings of the International conference on natural language processing (ICON 2010).* Kharagpur, India. http://www.cs.uccs.edu/~jkalita/work/reu/REUFinalPapers2010/Kaufmann.pdf

Kern, M. L., Eichstaedt, J. C., Schwartz, H. A., Dziurzynski, L., Ungar, L. H., Stillwell, D. J., … & Seligman, M. E. P. (2014a). The online social self: An open vocabulary approach to personality. *Assessment, 21,* 158-169. http://dx.doi.org/10.1177/1073191113514104

Kern, M. L., Eichstaedt, J. C., Schwartz, H. A., Park, G., Ungar, L. H., Stillwell, D. J., … & Seligman, M. E. P. (2014b). From “sooo excited!!!” to “so proud”: Using language to study development. *Developmental Psychology, 50,* 178-188. http://dx.doi.org/10.1037/a0035048

Kern, M., L. & Sap, M. (2015, February). *Do you feel what I feel? Cultural variations in linguistic expressions of emotion*. Symposium talk presented at the 16th annual meeting of the Society of Personality and Social Psychology, Long Beach, CA.

Kinsella, S., Murdock, V., & O’Hare, N. (2011). "I’m eating a sandwich in Glasgow": Modeling locations with tweets. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents, SMUC ’11* (pp. 61-68). Glasgow, Scotland. http://neilohare.com/papers/smuc2011_paper01.pdf

Kong, L., Schneider, N., Swayamdipta, S., Bhatia, A., Dyer, C., & Smith, N. A. (2014). A dependency parser for tweets. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, *4*. https://www.cs.cmu.edu/~nschneid/twparser.pdf

Kong, L., Liu, Z., and Huang, Y. 2014. Spot: Locating social media users based on social network context. In *Proceedings of the VLDB Endowment 7*(13). http://www.vldb.org/2014/program/papers/demo/p1154-liu.pdf

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

Kuhn, M. (2015). *Caret: Classification and regression training* [R package version 6.0-14]. https://github.com/topepo/caret/.

Lampos, V., & Cristianini, N. (2010). Tracking the flu pandemic by monitoring the social web. In Proceedings of the 2nd International Workshop on Cognitive Information Processing (pp. 411-416). http://dx.doi.org/10.1109/CIP.2010.5604088

Lancichinetti, A., Sirer, M. I., Wang, J. X., Acuna, D., Körding, K., & Amaral, L. A. N. (2015). High-reproducibility and high-accuracy method for automated topic classification. *Physical Review X*, *5*, 011007. http://dx.doi.org/10.1103/PhysRevX.5.011007

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. http://dx.doi.org.10.1037/0033-295X.104.2.211

Li, R., Wang, S., & Chang, K. C.-C. (2012). Multiple location profiling for users and relationships from social network and content. In *Proceedings of the VLDB Endowment, 5*, 1603-1614. http://vldb.org/pvldb/vol5/p1603_ruili_vldb2012.pdf

Li, W., & McCallum, A. (2006). Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the 23rd International Conference on Machine learning*. https://people.cs.umass.edu/~mccallum/papers/pam-icml06.pdf

Lian, X.-C., Li, Z., Wang, C., Lu, B.-L., & Zhang, L. (2010). Probabilistic models for supervised dictionary learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2305-2312). http://bcmi.sjtu.edu.cn/~lianxiaochen/papers/cvpr2010.pdf

liev, R., Dehghani, M., & Sagi, E. (2015). Automated text analysis in psychology: methods, applications, and future developments. *Language and Cognition, 7,* 265-290. http://dx.doi.org/0.1017/langcog.2014.30

Lin, D. (1998, August). *Extracting collocations from text corpora*. Poster presented at 1st Workshop on Computational Terminology, Montreal, Canada. http://dx.doi.org/10.1.1.11.7962

MacCallum, A. K. (2002). *MALLET: A machine learning for language toolkit*. http://mallet.cs.umass.edu

Mahmud, J., Nichols, J., & Drews, C. (2012). Where is this tweet from? Inferring home locations of Twitter users. In *Proceedings of Sixth International AAAI Conference on Weblogs and Social Media, ICWSM '12*. Dublin, Ireland. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4605/5045

Mairal, J., Bach, F., Ponce, J., Sapiro, G., & Zisserman, A. (2009). Supervised dictionary learning. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in neural information processing systems (NIPS 2009)* (pp. 1033-1040). http://www.di.ens.fr/willow/pdfs/nips08b.pdf

Martinsson, P. G., Rokhlin, V., & Tygert, M. (2011). A randomized algorithm for the decomposition of matrices. *Applied and Computational Harmonic Analysis, 30,* 47–68. http://dx.doi.org/10.1016/j.acha.2010.02.003

McGee, J., Caverlee, J. A., & Cheng, Z. (2013). Location prediction in social media based on tie strength. In *Proceedings of CIKM 2013* (pp. 459-468). http://dx.doi.org/10.1145/2505515.2505544

Meyer, E. T., & Schroeder, R. (2014). *Digital transformations of research.* Cambridge: MIT Press.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of the ICLR Workshop 2013*. http://arxiv.org/pdf/1301.3781.pdf

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS, 2013*. http://arxiv.org/pdf/1310.4546.pdf

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence, 29*, 436-465. http://arxiv.org/pdf/1308.6297.pdf

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR), 41*, 10. http://dx.doi.org/10.1145/1459352.1459355

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

Pashler, H. & Wagenmakers, E.-J. (2012). Editors introduction to the special section on replicability in psychological science: A crisis of confidence? *Psychological Science, 7*, 528-530. http://dx.doi.org/10.1177/1745691612465253

Paul, M. J., & Dredze, M. (2011). You are what you Tweet: Analyzing Twitter for public health. In *Proceedings of the 5th International AAAI Conference on Web and Social Media (ICWSM)*. (pp. 265-272). http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2880/3264

Pennebaker, J. W. (2011). *The secret life of pronouns: What our words say about us*. New York: Bloomsburg Press.

Pennebaker, J. W., & Francis, M. E. (1996). Cognitive, emotional, and language processes in disclosure. *Cognition and Emotion, 10*, 601-626. http://dx.doi.org/10.1080/026999396380079

Pennebaker, J. W., & Francis, M. E. (1999). *Linguistic Inquiry and Word Count: LIWC*. Mahwah, NJ: Erlbaum.

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology, 77*, 1296-1312. http://dx.doi.org/10.1037/0022-3514.77.6.1296

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

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for word representation. http://nlp.stanford.edu/projects/glove/

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *14*, 130-137.

Press, G. (2013, May). A very short history of big data [blogpost]. *Forbes*. Retrieved from http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/

Ritter, R. S., Preston, J. L., & Hernandez, I. (2014). Happy tweets: Christians are happier, more socially connected, and less analytical than atheists on Twitter. *Social Psychological and Personality Science, 5*, 243-249. http://dx.doi.org/10.1177/1948550613492345

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence* (pp. 487-494). http://mimno.infosci.cornell.edu/info6150/readings/398.pdf

Rout, D., Preotiuc-Pietro, D., Bontcheva, K., & Cohn, T. (2013). Where's wally?: A classification approach to geolocating users based on their social ties. In HT'13: *Proceedings of the 24th ACM Conference on Hypertext and Social Media* (pp. 11-20). http://dx.doi.org/10.1145/2481492.2481494

Rozenfeld, M. (2014, October). Your questions about big data answered. *IEEE Institute*. http://theinstitute.ieee.org/ieee-roundup/opinions/ieee-roundup/your-questions-about-big-data-answered

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. *Computational Linguistics and Intelligent Text Processing, 2276*, 1-15. http://dx.doi.org/10.1007/3-540-45715-1_1

Schütze, H., & Pedersen, J. O. (1997). A co occurrence-based thesaurus and two applications to information retrieval. *Information Processing & Management*, *33*, 307-318. http://dx.doi.org/10.1016/S0306-4573(96)00068-4

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Lucas, R. E., Agrawal, M., … & Ungar, L. H. (2013a). Characterizing geographic variation in well-being using tweets. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM)*. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6138/6398

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*, e73791. http://dx.doi.org/10.1371/journal.pone.0073791

Schwartz, H. A., Park, G. J., Sap, M., Weingarten, E., Eichstaedt, J. C., Kern, M. L., …, & Ungar, L. H. (2015). Extracting human temporal orientation from Facebook language. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL HLT 2015)*. http://www.seas.upenn.edu/~hansens/tempor-naacl15-cr.pdf

Shah, D. V., Cappella, J. N., & Neuman, W. R. (2015). Big data, digital media, and computational social science: Possibilities and perils. *Annals of the American Academy, 659*, 6-13. http://dx.doi.org/10.1177/0002716215572084

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420-428. http://dx.doi.org/10.1037/0033-2909.86.2.420

Sumner, C., Byers, A., & Shearing, M. (2011, December). *Determining personality traits and privacy concerns from Facebook activity.* Black Hat Briefings Conference, Abu Dhabi, United Arab Emirates.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association, 101*, 1566-1581. http://www.jstor.org/stable/27639773

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, *58*, 267-288. http://statweb.stanford.edu/~tibs/lasso/lasso.pdf

Turian, J. Ratinov, L., & Bengio, Y. (2010). Word representations: A simple and general method for semi-supervised learning. In ACL'10: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (pp. 384-394). http://www.aclweb.org/anthology/P10-1040

Vaillant, G. E. (2012). *Triumphs of experience: The men of the Harvard Grant Study*. Cambridge, MA: Belknap Press.

Wallach, H. M. (2006). Topic modeling: Beyond bag-of-words. In *Proceedings of the 23rd International Conference on Machine learning*. http://people.ee.duke.edu/~lcarin/icml2006.pdf

Wang, C., Thiesson, B., Meek, C., &Blei, D. (2009). Markov topic models. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS)*. http://www.cs.cmu.edu/~chongw/papers/WangThiessonMeekBlei2009.pdf

Weeg, C., Schwartz, H. A., Hill, S., Merchant, R. M., Arango, C., & Ungar, L. (2015). *Using Twitter to measure public discussion of diseases*. Unpublished manuscript.

Weinberg, C., & Gordon, A. S. (2015). *Insights on privacy and ethics from the web's most prolific storytellers*. The 7th Annual ACM Web Science Conference (WebSci '15), Oxford, UK. http://dx.doi.org/10.1145/2786451.2786474

White, R. W., Harpaz, R., Shah, N. H., DuMouchel, W., & Horvitz, E. (2014). Toward enhancing pharmacovigilance using patient-generated data on the internet. *Clinical Pharmacology & Theurapeutics*. http://dx.doi.org/10.1038/clpt.2014.77

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, 22-31. http://dx.doi.org/10.3758/BF03195494

Yang, S.-H., White, R. W., & Horvitz, E. (2013). Pursuing insights about healthcare utilization via geocoded search queries. In *Proceedings of 36th Annual ACM SIGIR Conference.* Dublin, Ireland. http://research.microsoft.com/en-us/um/people/horvitz/Geocoded_health_search_SIGIR2013.pdf

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*, 363-373. http://dx.doi.org/10.1016/j.jrp.2010.04.001

Yarkoni, T. (2012). Psychoinformatics: New horizons at the interface of the psychological and computing sciences. *Current Directions in Psychological Science, 21*, 391-397. http://dx.doi.org/10.1177/0963721412457362

Zhu, J., Ahmed, A., & Xing, E. P. (2012). MedLDA: Maximum margin supervised topic models. *Journal of Machine Learning Research, 13*, 2237-2278. http://www.jmlr.org/papers/volume13/zhu12a/zhu12a.pdf

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B (Statistical Methodology), 67*, 301-320. http://dx.doi.org/10.1111/j.1467-9868.2005.00503.x

Table 1

Examples of signal discrepancies in identifying words indicating positive emotion and meaning in life

Category	Sub-category	Description	Examples
Lexical ambiguity	Wrong part of speech (POS)	Term is the wrong part of speech, such as a verb instead of a noun	"My father will **tender** the company to me" (verb instead of a positive emotion noun) "I saw the movie **Happy** Feet" (proper noun instead of a noun)
	Wrong word sense (WS)	Term is used with a different meaning	"my muscles feel **tender**" (indicates soreness, not positive emotion)
Signal negation	Strict negation	Term is used with a clear negative qualifier or adverb, which negates the term	"I am **not happy**" "I **haven't** found a **purpose** for my life"
	Desiring	User is wishing for something, implying the opposite	"I **wish** I could be **happy**"
Weak or mixed signal	Conjunction of mixed	Term signals one category, but a conjunction qualifies it to suggest the opposite feeling	"My friends are **great, but** they really annoy me" (possibly ignore signal in the first clause)
	Reasoning against	A term is used that reasons against an idea	"storing up wealth to hand it over to others. This too is **meaningless**, a chasing after the wind"
Duplicated collocations	Internet meme	Duplicated text that spreads across users	"This is cancer awareness month. Put this up for 1 hour if you **love** someone who has or had cancer. I was **proud** to. Will you?"
	Quote	Clearly part of a quote	"As Anne Frank said, 'whoever is **happy** will make others **happy** too'"
	Other collocations	Catch all category for other common sequences of words	"**Merry** Christmas" "**Good** evening"

Table 2

Closed vocabulary analysis example: Frequency of LIWC social, affective, and cognitive processes categories across 72,709 users, and correlations with self-rated extraversion

LIWC category	Sample words	N	Mean	SD	Min	Max	β_e
Social processes	Buddies*, love, somebod*,listen, talked	72709	0.068	0.021	0.000	0.198	.04
Family	Brother*, cousin*, mum, sis, relatives	72709	0.004	0.003	0.000	0.069	.03
Friends	Acquainta*, bf*, guest*, pal, colleague	72709	0.002	0.002	0.000	0.029	.05
Humans	Child, citizen, person, societ*, members	72709	0.007	0.003	0.000	0.044	.06
Affective processes	Discomfort*, trouble*, ugh, miss, grin	72709	0.065	0.015	0.002	0.188	.07
Positive emotion	Hope, happy, joy*, okay, fabulous*	72709	0.045	0.013	0.000	0.184	.13
Negative emotion	Distrust*, lost, tense*, mad, grief	72709	0.020	0.008	0.000	0.095	-.07
Anxiety	Obsess*, rigid*, shaky, tense*, scare*	72709	0.002	0.001	0.000	0.031	-.04
Anger	Rage*, frustrate*, fuming, temper, hostil*	72709	0.008	0.005	0.000	0.085	-.05
Sadness	Pity*, remorse, sorrow*, weep*, low*	72709	0.004	0.002	0.000	0.067	-.04
Cognitive processes	Anyhow, directly, true, suppose, based	72709	0.110	0.026	0.000	0.217	-.05
Causation	Foundation*, made, allow*, caus*, control*	72709	0.010	0.003	0.000	0.048	-.06
Certainty	Absolutely, clear, definite, fact, never	72709	0.011	0.004	0.000	0.081	*ns*
Discrepancy	Needs, should, want, could, mustn't	72709	0.013	0.005	0.000	0.054	-.05
Exclusive	But, not, or, versus, without	72709	0.019	0.006	0.000	0.057	-.07
Inclusive	Add, came, open, out, with	72709	0.030	0.010	0.000	0.110	.04
Inhibition	Bans, brake*, cease*, distinct*, guard*	72709	0.004	0.002	0.000	0.065	*ns*

LIWC category	Sample words	N	Mean	SD	Min	Max	β_e
Insight	Accept, learn*, notice*, choice*, prefer*	72709	0.014	0.005	0.000	0.045	-.09
Tentative	Almost, change, depend, dunno, partly	72709	0.018	0.006	0.000	0.059	-.08

Note. LIWC is constructed hierarchically, such that categories (e.g., positive emotion) are nested within larger categories (e.g., affect). Sample words were randomly selected from the LIWC dictionaries (Pennebaker & Francis, 1999). β_e = correlations between each category and self-rated extraversion scores, controlled for age and gender. ns = non-significant, * indicate wildcards, which capture variants of the word stem.

Table 3

Top ten words for topics with "happy" and "play", across 50, 500, and 2000 topics generated from the MyPersonality dataset

Generated	Occurrences	Top 10 words comprising each topic
		HAPPY
50	2	happy, christmas, year, family, friends, hope, merry, thanksgiving, wonderful, easter
		happy, birthday, day, love, wishes, mom, miss, wonderful, dad, family
500	8	day, happy, mothers, mother's, mom, mother, wonderful, moms, mommy, mama
		day, happy, valentines, fathers, valentine's, father's, dad, independence, dads, single
		birthday, happy, wishes, bday, wished, b-day, birthdays, present, celebrate, cake
		year, happy, 2010, 2011, joy, wishing, bring, happiness, safe, diwali
		happy, 4th, july, halloween, year, fireworks, safe, fourth, holiday, holidays
		happy, thanksgiving, easter, family, thankful, turkey, holiday, bunny, enjoy, eggs
		happy, birthday, anniversary, wishing, brother, son, bday, daddy, mommy, celebrate
		happy, makes, sooo, soo, soooo, easter, thanksgiving, camper, ending, sooooo
2000	20	happy, birthday, mommy, daddy, mama, momma, dearest, bestest, 21st, 18th
		happy, birthday, sis, lil, bday, b-day, luv, cousin, 21st, nephew
		happy, mothers, mother's, mom, moms, mother, mommy, mom's, mama, mommies
		happy, makes, camper, unhappy, extremely, happier, smiling, satisfied, contented, content
		happy, diwali, wishing, eid, happiness, mubarak, holi, festival, prosperous, gibran
		easter, happy, bunny, eggs, egg, hunt, holidays, risen, candy, basket
		happy, birthday, brother, wishing, 18th, 21st, xxxx, 16th, monthsary, nephew
		year, happy, 2010, 2011, chinese, 2009, cheers, prosperous, tiger, rabbit
		happy, independence, friendship, valentines, canada, valentine's, republic, memorial, australia, boxing
		year, happy, joy, happiness, bring, 2010, 2011, health, wishing, brings
		happy, fathers, father's, dad, dads, father, daddy, dad's, mothers, papa
		4th, july, happy, fireworks, fourth, safe, independence, bbq, 5th, quarter
		happy, birthday, celebrate, anniversary, celebrating, birthdays, dad's, b-day, b'day, mom's
		happy, valentines, valentine's, single, valentine, hump, pi, awareness, singles, v-day
		happy, birthday, grandma, mama, aunt, beth, mary, anniversary, papa, grandpa
		birthday, happy, wishes, wished, 21st, 18th, bithday, happpy, meeee, birthday's

Generated	Occurrences	Top 10 words comprising each topic
		birthday, happy, wishes, wished, birthdays, thankyou, birthday's, individually, 11:11, manthy
		happy, thanksgiving, halloween, holidays, easter, sabbath, birthdays, 420, festivus, fiesta
		hasn't, yesterday, happened, arrived, started, choose, unhappy, marx, events, groucho
		happy, thanksgiving, turkey, thankful, gobble, holiday, feast, parade, turkeys, meal
		PLAY
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
		place, chuck, find, meet, play, birth, norris, interesting, babies, profile
		play, playing, game, games, xbox, halo, wii, video, mario, 360
		play, playing, game, ball, games, played, golf, tennis, poker, cards
2000	25	golf, played, ultimate, frisbee, mini, ball, balls, golfing, tennis, disc
		play, game, let's, role, sims, rules, chess, basketball, plays, poker
		words, comment, note, play, wake, jail, copy, paste, sport, fair
		black, cod, ops, playing, play, mw2, modern, warfare, ps3, online
		game, team, won, win, played, boys, soccer, season, proud, football
		soccer, football, game, play, team, basketball, playing, ball, practice, field
		kids, park, playing, boys, played, pool, blast, playground, swimming, toys
		sand, beach, water, toes, carl, grain, playin, mountain, rocks, desert
		music, band, playing, piano, guitar, songs, sound, metal, bass, played
		na, stuck, everyday, ki, replay, melody, ami, er, ta, singin
		http://www.youtube.com, feature, related, =p, marcus, channel, double, user, nr, youtube_gdata_player
		guitar, bass, drum, playing, amp, drums, string, strings, electric, acoustic
		play, guitar, learn, piano, learning, playing, learned, lessons, songs, rules
		games, play, playing, game, video, played, card, board, begin, playin
		play, playing, starcraft, warcraft, sims, ii, beta, online, nerds, nerd
		watchin, sittin, chillin, waitin, doin, havin, gettin, eatin, playin, drinkin
		pokemon, playing, mon, shiny, version, pikachu, pok, cards, ds, ash
		player, dvd, cd, record, printer, bought, set, mp3, ink, borrow

Generated	Occurrences	Top 10 words comprising each topic
		anime, manga, naruto, bleach, episode, series, cosplay, episodes, alchemist, japanese
		xbox, 360, play, ps3, playing, games, creed, assassin's, playstation, assassins
		hero, guitar, playing, rockband, dj, devin, playin, beatles, expert, metallica
		didn't, eat, parents, survived, kid, played, exist, bike, telling, raised
		mario, wii, playing, super, games, nintendo, zelda, bros, fit, ds
		play, playing, tennis, cards, wii, played, poker, ball, basketball, pool
		won, team, poker, win, tournament, league, competition, played, winning, champion

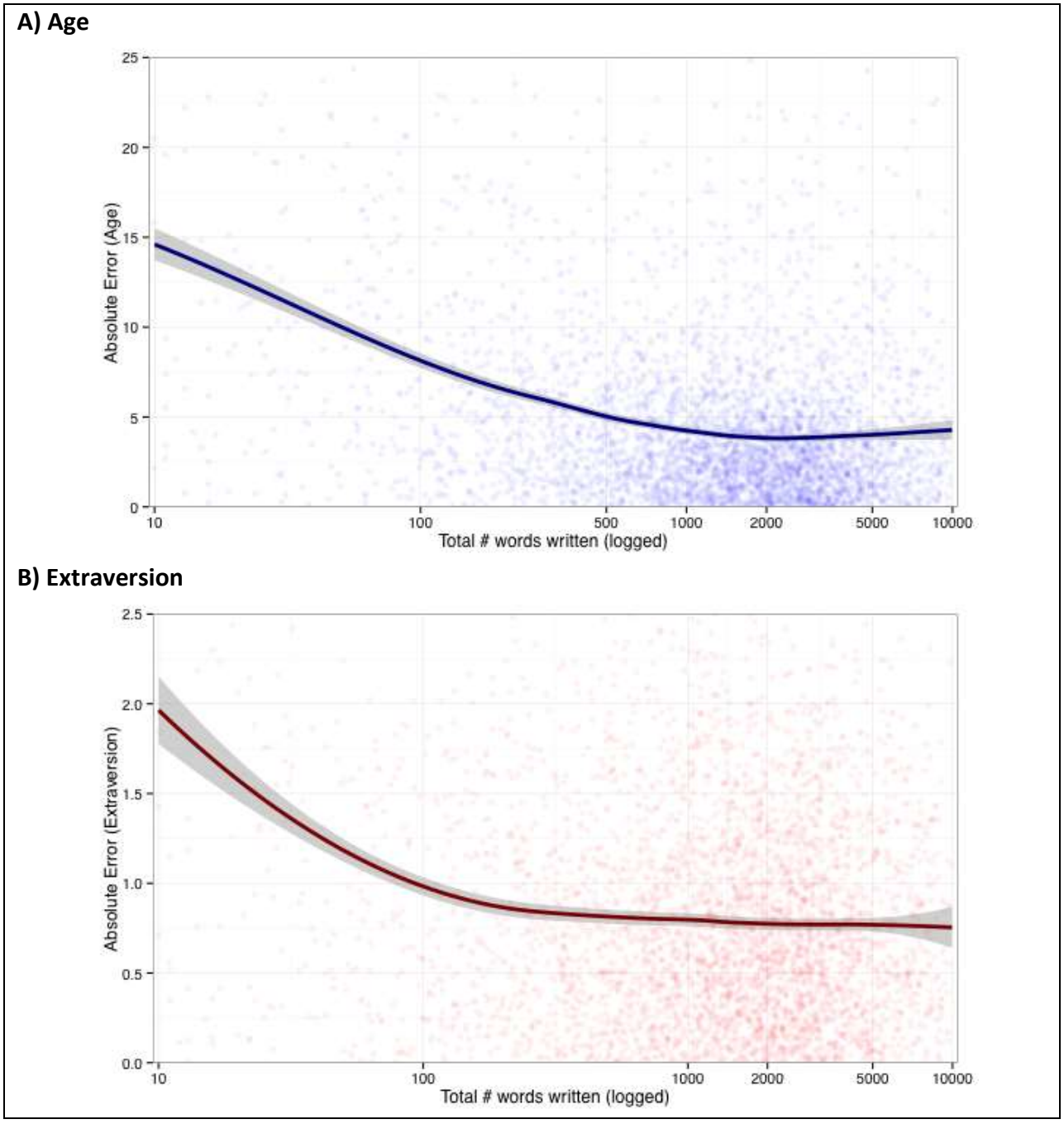
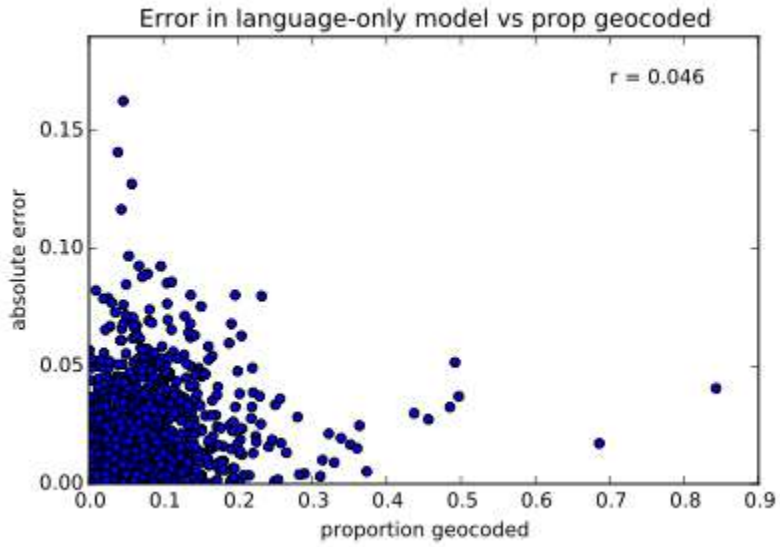


Figure 1. Impact of word count estimates of age (top) and extraversion (bottom) on accuracy, based on 4,000 randomly selected users of the MyPersonality dataset. The y-axis is the mean absolute error of the model (i.e., the average absolute value of the difference between a person's age or extraversion score predicted from their words) and the x-axis is total words written (logarithmically scaled). The errors are in years for age and in normalized scores for extraversion, so the units are not directly comparable. The line on each graph was fit with LOESS regression (Cleveland, 1979) and the shaded area indicates the 95% confidence interval.

A) Predicting life satisfaction, uncontrolled model



B) Predicting life satisfaction, controlling for demographics

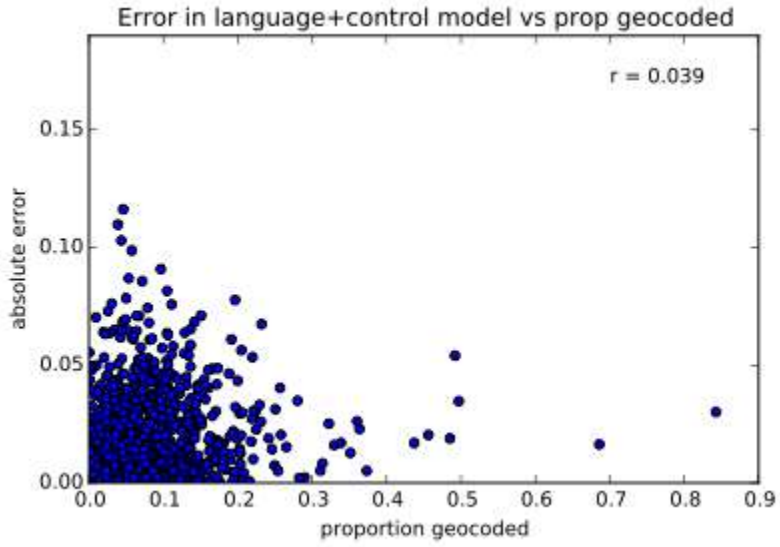


Figure 2. Error of using the Twitter free response location as an estimate for county location versus using geocodes to predict county level life satisfaction from words, phrases, and topics (*N* = 1,071 counties, 148 million tweets) in the uncontrolled model (top) and controlling for demographics (bottom).

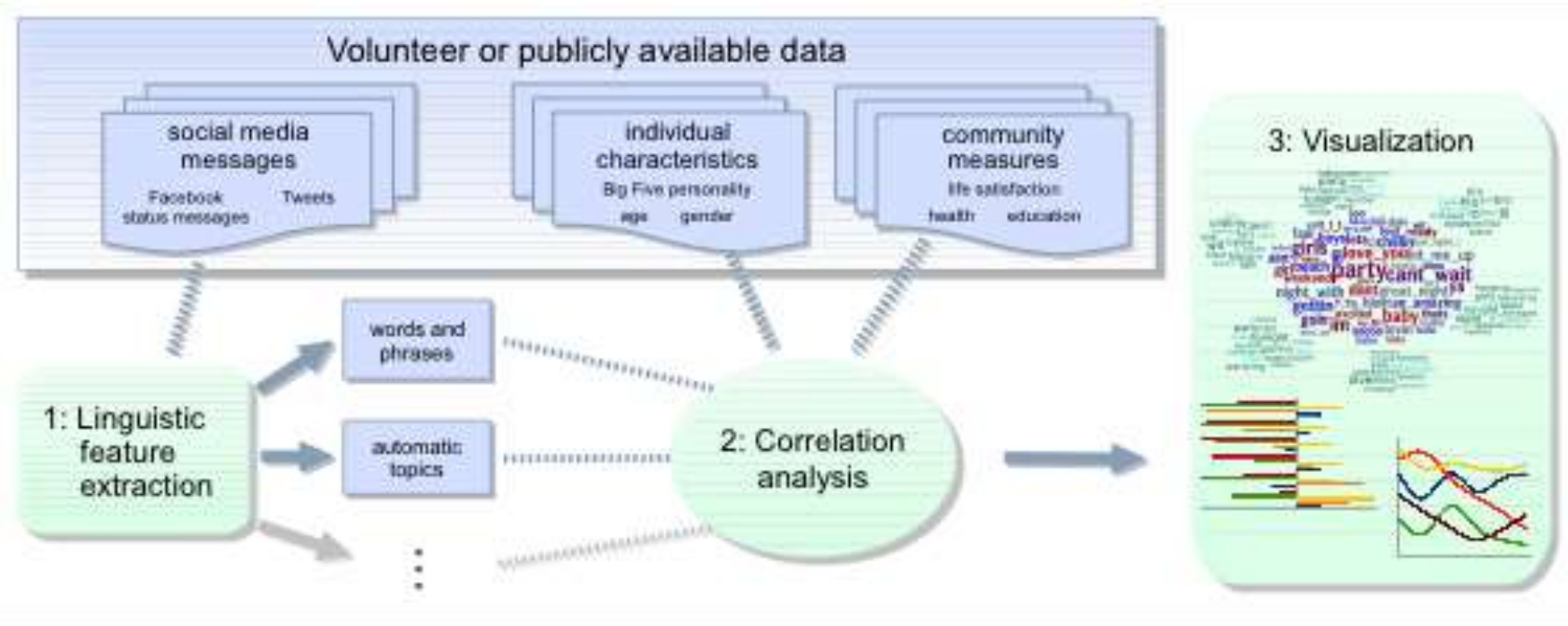


Figure 3. Illustration of the differential language analysis (DLA) process. Words, phrases, topics, and other linguistic features are extracted from social media messages (Step 1). The relative frequencies of those features are correlated with other characteristics (Step 2). Results are visualized to aid interpretation (Step 3). Illustrations might include word clouds of correlated words and phrases (part 3, top center), word cloud topics (part 3, top), frequency distributions (part 3, bottom left) and loess lines for patterns across time or age (part 3, bottom right). Figure adapted from Schwartz et al., 2013b

A) High extraversion

B) Low extraversion

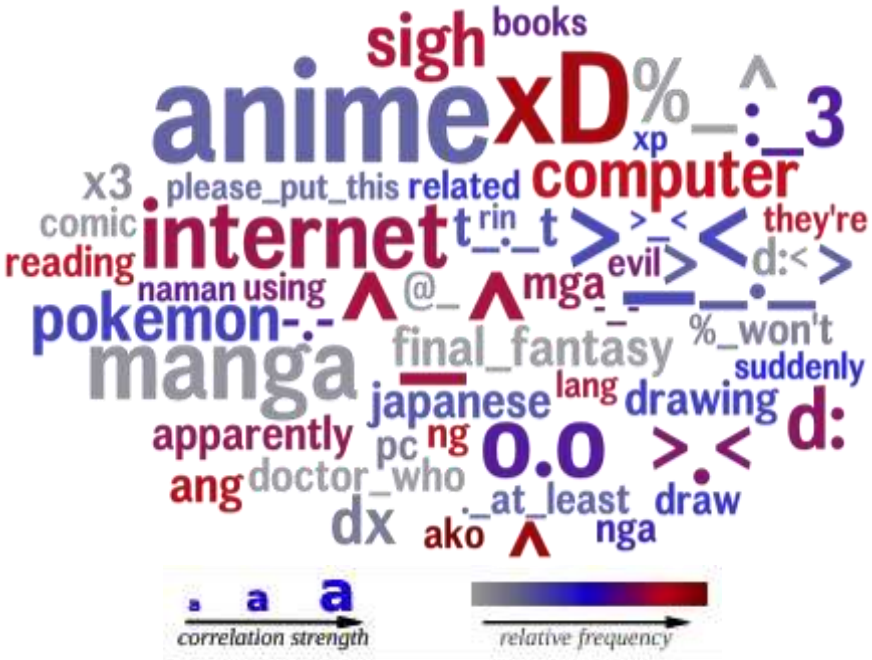


Figure 4. Example of DLA: words and phrases that were most strongly positively (top) and negatively (bottom) correlated with extraversion, across 70,000 users. The size of the word indicates the correlation with extraversion (larger = stronger correlation), and color indicates frequency (grey = infrequent, blue = moderate frequency, red = frequently used). Figure adapted from Schwartz et al., 2013b.

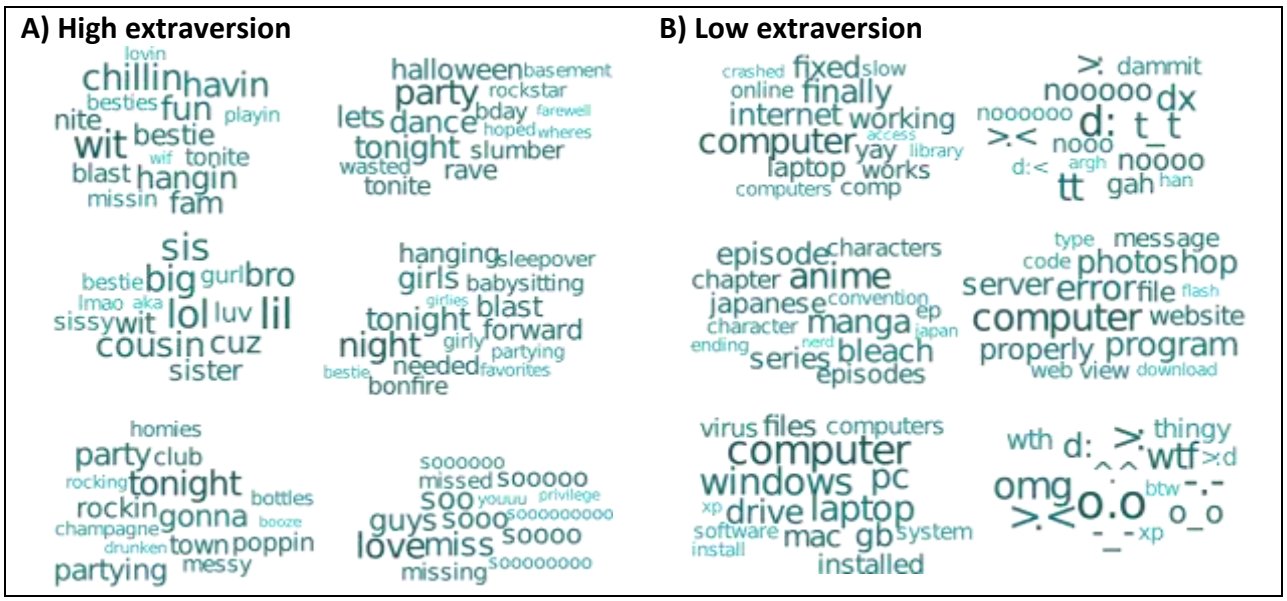


Figure 5. Example of automatically created topics, illustrating the topics most strongly positively (top) and negatively (right) correlated with extraversion across 70,000 users. The size of the word indicates its weight within the topic, such that the larger the word, the more it represents that topic. Figure adapted from Schwartz et al., 2013b.