THE GERONTOLOGICAL SOCIETY OF AMERICA®

Journals of Gerontology: Psychological Sciences
cite as: *J Gerontol B Psychol Sci Soc Sci*, 2017, Vol. 72, No. 5, 742–751
doi:10.1093/geronb/gbx014
Advance Access publication March 2, 2017

OXFORD

Special Issue: Methodological Innovations in Gerontology: Advances in Psychosocial Research: Special Article

Machine Learning, Sentiment Analysis, and Tweets: An Examination of Alzheimer's Disease Stigma on Twitter

Nels Oscar[1], Pamela A. Fox[2], Racheal Croucher[2], Riana Wernick[3], Jessica Keune[4], and Karen Hooker[2]

[1]School of Electrical Engineering and Computer Science, Oregon State University, Corvallis. [2]School of Social and Behavioral Health Sciences, Oregon State University, Corvallis. [3]Department of Integrative Biology, Oregon State University, Corvallis. [4]School of Biological and Population Health Sciences, Oregon State University, Corvallis.

Correspondence should be addressed to Nels Oscar, BS, Kelley Engineering Center, 2500 NW Monroe Avenue, Corvallis, OR 97331. E-mail: oscarn@oregonstate.edu.

Decision Editor: Shevaun Neupert, PhD

Received July 6, 2016; Editorial Decision Date January 19, 2017

Abstract

Objectives: Social scientists need practical methods for harnessing large, publicly available datasets that inform the social context of aging. We describe our development of a semi-automated text coding method and use a content analysis of Alzheimer's disease (AD) and dementia portrayal on Twitter to demonstrate its use. The approach improves feasibility of examining large publicly available datasets.
Method: Machine learning techniques modeled stigmatization expressed in 31,150 AD-related tweets collected via Twitter's search API based on 9 AD-related keywords. Two researchers manually coded 311 random tweets on 6 dimensions. This input from 1% of the dataset was used to train a classifier against the tweet text and code the remaining 99% of the dataset.
Results: Our automated process identified that 21.13% of the AD-related tweets used AD-related keywords to perpetuate public stigma, which could impact stereotypes and negative expectations for individuals with the disease and increase "excess disability".
Discussion: This technique could be applied to questions in social gerontology related to how social media outlets reflect and shape attitudes bearing on other developmental outcomes. Recommendations for the collection and analysis of large Twitter datasets are discussed.

Keywords: Attitudes—Data mining—Social media—Stigma

Given that aging is both a biological process and a social construct, gerontologists care a great deal about the social context of aging. Rapid proliferation of social media has shaped the social environment in recent years. The advent of Twitter in 2006 has popularized microblogging, which now provides its 320 million users the ability to post comments and status updates of up to 140 characters (Isaac, 2016). As a result, close to real time comments and updates (i.e., tweets) have become commonplace. This is important considering that tweets are often treated as information sources and are cited in traditional information outlets such as news media (Kwak, Lee, Park, & Moon, 2010). The entirety of the public tweet record is even being recorded and preserved in the national archives. However, despite the availability of this (and other) large datasets and some existing methodological tools for data mining, use of these data by psychologists has tended to lag behind other disciplines—and this is especially true in social gerontology. Improvements on existing methodologies are needed to enable social science researchers to make better use of

these valuable data to understand sentiment about aging-related issues.

We believe that the combination of machine learning (ML) algorithms and existing content analysis tools can allow researchers to address problems particular to the field of gerontology. There are a number of existing tools designed to examine tweets for sentiment, specific mentions, and other attributes. Unfortunately, the most user friendly of these tools have often been developed to target general aspects of tweets, such as general positive/negative sentiment—as in Microsoft's NLP Toolkit. However, none of the existing sentiment analysis tools are designed to capture this stigma. In this article, we develop and demonstrate a supervised method for coding the content of sample of tweets on several dimensions relevant to Alzheimer's disease (AD) stigma.

Twitter Analysis Techniques

To date, Twitter research has been conducted across a variety of disciplines including economics, biology, computer science, engineering, and medicine (Williams, Terras, & Warwick, 2013a). Psychological analyses were slow to emerge as of 2012 (Zimmer & Proferes, 2014), but experienced marked uptick since then. Empirical reports using Twitter data have been organized according to their aims, and aspects of tweets measured, using the nonexclusive categories: content analysis, sentiment analysis, event detection, user studies, prediction, and GIS analysis (Zimmer & Proferes, 2014). Content analysis, and more specifically, sentiment analysis are closely related methods of text analysis. They have featured prominently in the existing social science studies (Zimmer & Proferes, 2014) where positivity of the content is frequently the dimension of focus.

Content analysis uses the text of tweets as a basis for detecting themes. For instance, the presence of words related to medical conditions has been used to detect the presence of diseases and predict their spread (Signorini, Segre, & Polgreen, 2011; Williams, Terras, & Warwick, 2013b). Sentiment analysis is a form of content analysis specifically aimed at describing the affective or emotional tone present in text (Pang & Lee, 2008) based on psychological evidence about the emotional meaning of the constituent words or phrases (e.g., Anderson, 1968; Tausczik & Pennebaker, 2010). Sentiment analysis focuses on the positivity of the content and is the primary technique by which public opinion is gauged using Twitter data, and has been used to track response to the Boston Marathon bombing (Cassa, Chunara, Mandl, & Brownstein, 2013), predict short term fluctuation in the stock market (Bollen, Mao, & Zeng, 2011), describe attitudes toward political candidates (Tumasjan, Sprenger, Sandner, & Welpe, 2010) and consumer products (Jansen and Zhang, 2009), forecast election outcomes (e.g., Chung & Mustafaraj, 2011), and detect language associated with depressive symptoms (De Choudhury, Counts, & Horvitz, 2013; Park, Cha, & Cha, 2012).

A number of sentiment analysis tools are available, and while they all share in common the basic aim of quantifying affective dimensions of text, they differ in the process by which this is achieved. The distinction between lexicon-based and machine-learning based approaches is relevant for our purposes. A lexicon-based approach utilizes a "dictionary" of words with known affective meaning (i.e., an associated positivity score) to detect the presence of affective language in a text sample, and generate a rating (e.g., a positivity and negativity rating; Zhang, Ghosh, Dekhil, Hsu, & Liu, 2015). For instance, the popular Linguistic Inquiry and Word Count (LIWC; Pennebaker, Booth, & Francis, 2007) uses a broad lexicon to generate frequencies of words that represent a number of dimensions such as degree of overall emotionality, positivity, and negativity for the purpose of sentiment analysis. Word counts can be generated according to the established lexicons for other dimensions such as social relationships, biological processes, and the life domains of family, health, work, and leisure. LIWC could be used with Twitter data, for example, to generate a score for each tweet on the dimensions of positivity or negativity, though there are some concerns about applying LIWC to such short texts. Other lexicon-based tools have been created specifically for the analysis of microblogs. Affective Norms for English Words list (ANEW; Nielsen, 2011), is a lexicon-based approach that uses a word list specifically attuned for Twitter in that it includes slang, abbreviation and other conventions unique to the Twitter format. However, the word list is relatively small (about 4,000 words) and, like most sentiment analysis tools, reports only on a small set of general dimensions: affective valence, arousal (i.e., calm to excited), and dominance. Although lexicon-based approaches are relatively easy to use, limitations of the approach include low recall, which occurs when a lexicon is too small to capture a sufficient portion of the text sample, or when the presence of lexicon words is low in the text. An additional limitation is that researchers may be interested in more nuanced dimensions beyond general positivity and negativity, or the measures available for entries in the lexicon. Beyond lexicon-based approaches, some sentiment analysis techniques use ML approaches.

ML Tools

Beyond generating word counts, ML-based tools for text analysis provide a probabilistic approach to classifying text according to desired dimensions (i.e., positivity) in a way that allows learning algorithms to improve in accuracy and make determinations about new text samples (Mitchell, 1997; Hall, Witten, & Frank, 2011). A classifier is constructed based on the input provided (i.e., sample texts known to represent positive or negative sentiment, referred to as "training data"). As with the lexicon-based content analysis tools described in the previous paragraph, a ML-based sentiment analysis may specify attributes of the target dimension using

a lexicon or similar established system for "tagging" characteristics of language (e.g., Treetagger; Schmid, 1994). The difference is that with increased training data, the ML-based method adjusts the weights of the individual attributes in the mathematical function generating the probability that a text sample belongs to the target class (i.e., that the text sample represents positive sentiment).

Some ML-based sentiment analysis tools include Microsoft NLP Toolkit, the WEKA data mining toolkit (Hall et al., 2011), SciKitLearn, and others. Although these tools can improve in accuracy based on feedback and increased training data input, they are limited in that the dimensions assessed are positivity and negativity. Some researchers have endeavored to capture more complex dimensions of language, such as sarcasm, but have concluded that "lexical features alone are not sufficient for identifying sarcasm and that pragmatic and contextual features merit further study" (Gonzalez-Ibanez, 2011). As a result, we expect that our dimensions of interest will vary in their identifiability using ML-based sentiment analysis.

A Sample Investigation of ML-Based Sentiment/Content Analysis

Our consideration of ML tools and the existing sentiment/content analysis tools lead us to identify our central aim of better understanding how we can apply ML to sentiment analysis to improve efficiency in analysis of large datasets. Our primary methodological question of interest, and the focus of this article, was: Can a ML algorithm use a relatively small amount of input information provided by researchers to automate the coding process for tweet content on six dimensions with reasonable accuracy and reliability? To address this, we developed a semi-automated coding method and demonstrate its use in an example study analyzing the content of posts related to AD and dementia on Twitter. Negative attitudes toward AD and dementia, the recent focus of the 2012 World Alzheimer's Report (Batsch & Mittelman, 2012), can be harmful to the stigmatized individuals when the resulting shame, guilt, hopelessness, and social exclusion, lead to delayed diagnosis (Mukadam & Livingston, 2012), inability to cope, decreased quality of life (Burgener, Buckwalter, Perkhounkova, & Liu, 2015) and increased burden of dementia (e.g., excess disability, Sabat, 2001). Stigma also affects friends, family, and caregivers of individuals with dementia when these close others become the target of stigmatizing views by association (stigma by association or "courtesy stigma;" Werner & Heinik, 2008). Given these negative consequences in combination with demographic trends of population aging projecting a threefold increase in the number of individuals with dementia worldwide from 43 million today to more than 131.5 million by 2050 (Batsch & Mittelman, 2012), it is surprising that little is known about the prevalence of public stigma in popular social media outlets. Thus, the research questions in our sample investigation of AD stigma were a) what is the prevalence of public stigma related to AD in a large sample of tweets posted by a broad sample of English speakers? and b) what are the sources of AD related tweets (private user versus organization)? We expected the results to confirm the feasibility and validity for using this semi-automated coding to establish evidence in support our hypothesis that Twitter is being used by a substantial subset of users to perpetuate AD stigma. We expected a substantial proportion of information-related tweets (i.e., links to articles, websites, resources about AD) to be posted by organizations while individual users are more likely to generate tweets based on personal experience. Additionally, we expected that tweets by organizations (compared to individuals) would demonstrate a lesser degree of stigma. And finally, we use a comparison of our results with LIWC analysis of our data to provide a check on our manual coding procedures and demonstrate the value of our ML approach compared to a lexicon-based sentiment analysis approach.

Methods

This study was designated as nonhuman subjects research by the Institutional Review Board at the researchers' institution.

Data Collection

Seventy-seven thousand eight publicly available tweets were collected continuously for 10 days in early 2014 using the Twitter keyword search application programming interface (API; "*API Overview*", 2016). These queries were subject to the overall volume limitations that Twitter imposes on their search APIs, meaning that we had access to only a portion of all tweets posted during our data collection window. Our research team identified a set of keywords related to AD, dementia, and cognitive decline as a basis for collecting tweets of interest. Tweets were restricted to English language Twitter accounts. For each tweet, we obtained data on the date, time, the user's publicly displayed name (their user "handle"), tweet body text including hashtags and links, and emoji. Account user names were not included.

Data Cleaning Prior to Manual Coding

To obtain an appropriate subsample for analysis, retweets ($N = 28{,}746$) were removed, as were tweets that had no clear relevance to AD or dementia (e.g., our keyword "cognitive" was too general and captured tweets that were not related to AD or dementia, so these $N = 17{,}112$ were removed). The 9 retained keywords that defined our sample for analysis were: "alz," "alzheimer," "dementia," "demented," "cognitive," "oldtimers," "memory loss," "senile," and "senility." Last, we removed posts by users whose handle included a keyword (e.g., senile_old_coot), but whose tweets had no content related to dementia. This resulted in a subsample of 31,150 tweets used for the analysis. User handles and links were included in the text sample.

Manual Coding Procedures

Two researchers manually coded a random subset ($N = 311$) of tweets. The results of the manual coding were assessed for satisfactory inter-rater reliability (a priori criterion > .61; Hallgren, 2012) then used as input for a ML algorithm that was used to automatically code the remaining tweets.

The two raters assessed tweets according to six dimensions, which were adapted from four dimensions in McNeil, Brna, and Gordon's (2012) examination of epilepsy stigma (metaphorical, personal experience, informative, joke/ridicule). We separated joke and ridicule into separate dimensions and added a dimension indicating whether the user was an individual or an organization (e.g., Alzheimer's Association, or a news agency). Instead of McNeil et al.'s original dichotomous rating scale we used a 5-point Likert scale in an attempt to capture greater variability (e.g., "Is this tweet speaking metaphorically?" 0 = *no*, 1 = *somewhat*, 2 = *fairly*, 3 = *significantly*, 4 = *completely*). Raters used a web interface that we created (Figure 1). Rater coding criteria are shown in Appendix A and source code for the creation of the web-based interface are provided (https://github.com/nels-o/ad_stigma). Sample tweets are shown in Figures 2 and 3.

Model Selection for ML Algorithm

The goal of the model was to predict, based on the manual coding input (i.e., training data), whether a new tweet contained ridicule, or any of the other five dimensions, given the content of the new tweet. Training our classifier (Breiman, 1996) involved the two steps of feature extraction and accuracy assessment (defined below). We used different classifiers for each of our six dimensions of interest. To parameterize our classifiers, we systemically examined the parameter space of the ML algorithms using grid search with threefold cross validation (Bergstra, Bardenet, Bengio, & Kegl, 2011).

Training the Classifier

Feature extraction is the process by which we determined what attributes such as *n*-grams or such as part of speech distribution information would be used to train the classifier to recognize instances of the six target dimensions: informative, joke, metaphorical, organization, personal experience, ridicule. *N*-grams refer to arrangements of words, for instance, a unigram is a single word (e.g., "happy"), a bi-gram is an arrangement of two words (e.g., "so sad"), and a tri-gram is

Figure 2. Example of a tweet that is informative (A) and a tweet that is a joke and ridiculing (B).

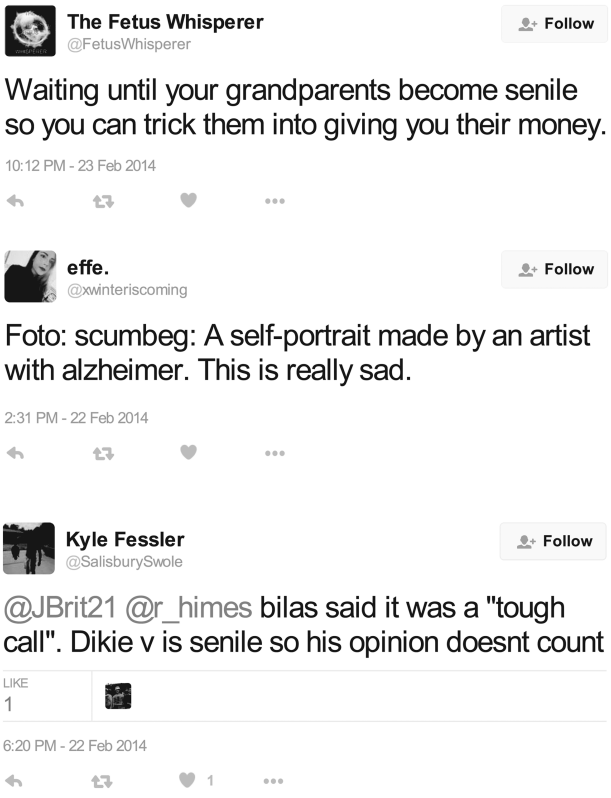

Figure 3. Examples of tweets classified as ridicule.

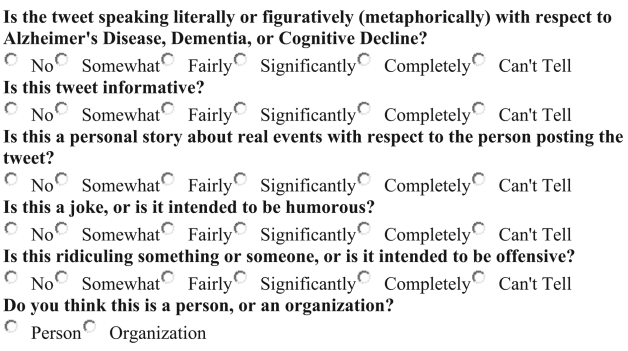

Figure 1. Example of web interface used in manual coding.

an arrangement of three words. Higher order *n*-grams are often necessary to detect negations in the meaning of text because the words "no" or "not" are linked to the thing they negate (e.g., Wilson, Gosling, & Graham, 2012). The features used by our classifiers were *n*-gram based; the presence of a given *n*-gram was a binary indicator of that feature. Our analysis included up to 5 g, and the "best" range of *n*-grams for the classifier was selected via grid search. We used a linear estimator to determine the estimated importance of each feature (i.e., *n*-gram) and features with a weight below the mean were discarded from the models.

Accuracy

Several procedures improved and assessed the accuracy of our classifier. We used 50 randomized trials of cross validation to evaluate the precision and recall of the classifiers. The classifier is sensitive to the population of tweets used to train it, so shuffling the tweets before threefold cross validation allow us to account for this. Testing accuracy was compared to the accuracy of a simple majority classifier (SMC) as defined and discussed below. Testing refers to the accuracy of machine-coded tweets compared to a subset of manually coded tweets not used as ML input.

Linguistic Inquiry and Word Count (LIWC)

After the manual coding was completed and ML model results obtained, the content of the 31,150 tweets was also analyzed using LIWC (both the manually coded subset and the automated coded entire dataset). LIWC provides information on a variety of dimensions such as prevalence of function words (pronouns), sentiment, punctuation, and a variety of content domains (e.g., family, health, work). We used a small number of dimensions from the LIWC output as a reference point to help validate manual coding and ML automated coding. LIWC is not equipped to detect linguistic indicators of stigma specifically, it does provide data on the percentage of words in each tweet that represent conceptually distinct, yet related measures of personal pronouns, positive emotion, and negative emotion. We expected, our "personal experience," dimension to positively correlate with the LIWC measure of personal pronouns. We also expected our dimension "ridicule" to be negatively associated with positive emotion and positively associated with negative emotion.

Results

Manual Coding

Manual coding was used as the input for the ML algorithm driving the automated coding process for the remainder of the dataset. Although manual ratings were collected using a 5-point Likert scale, for the analysis the ratings were collapsed into a dichotomous rating that reflected the presence of each dimension (i.e., for presence of ridicule 0 = *no*, 1 = *yes*). This was done so that we could use a binary classifier at the automated tagging stage. If a discrepancy between raters occurred, a tweet was coded 0 as not representing that dimension. This is a conservative strategy that may result in under-reporting the prevalence of the dimensions of interest, but we see it as preferable to introducing false positives. Manual coding inter-rater reliability was acceptable: informative (0.73), joke (0.57), metaphorical (0.73), organization (0.81), personal experience (0.47), and ridicule (0.73).

Manual coding results (N = 311) also provided evidence of AD stigma in the tweets: 43.41% informative, 23.79% joke, 21.22% metaphorical, 19.29% organization, 18.33% personal experience, and 24.50% ridicule (Table 3).

Automated Coding

We used the manual coding as input to train a classifier against the tweet text for each of the six dimensions to code the rest of the dataset (Table 1). This input was used to train the ML classifiers. To evaluate the classifiers, we used a subset of the tagged tweets as validation. Validation refers to the test accuracy of machine-coded tweets compared to a subset of manually coded tweets not used as ML input, and in our analysis testing accuracy ranged from 95.15% ("informative") to 86.38% ("organization"). Testing accuracy can be interpreted in comparison to the accuracy of a SMC, defined as the success rate of correct classification if each tweet were coded to correspond

Table 1. Manual Coding and Machine Learning Accuracy Results Across the Six Stigma Dimensions

	Manual ICC (%)	Testing accuracy (%)	Range (%)	*SD* (%)	SMC (%)	Diff (%)
Informative	80.78	95.15	91.35–99.04	1.82	56.59	38.56
Joke	65.68	87.42	80.77–92.31	2.79	76.21	11.21
Metaphorical	63.14	92.71	88.46–95.19	1.72	78.78	13.93
Organization	83.29	86.38	80.77–91.35	2.63	80.71	5.67
Personal	72.37	89.67	84.62–95.19	2.66	81.67	8.00
Ridicule	63.49	90.27	83.65–96.15	2.31	75.50	14.77

Notes: Manual ICC is the intraclass correlation (inter-rater reliability) for the 311 manually coded tweets. Testing refers to the accuracy of machine-coded tweets compared to a subset of manually coded tweets not used as machine learning input. Range and SD are reported for testing accuracy over 50 trials of threefold cross validation. This provides a measure of the stability of the model. SMC is the accuracy of a simple majority classifier and is used as a comparison for the more sophisticated classifiers. Diff = (testing accuracy – SMC). The "majority" proportions are based on the manually coded tweets.

to the class that makes up the majority of the manually coded tweets. For instance, a slight majority of the manually coded tweets were identified as not informative by the raters. A SMC based on these ratings would always predict that tweets are not informative, and it would be correct 56.59% of the time. We use this as a conceptual validation of our ML classifier to address whether the classifier performed better than guessing (Table 1). A greater difference between testing accuracy and SMC is desirable because it represents greater accuracy of our machine classifier. Differences ranged from 5.67% for "organization" to 38.56% for "informative."

Stigma in the Content of Tweets

21.13% of all tweets (N = 6,583) used AD-related keywords in a stigmatizing fashion. Prevalence of the six dimensions among all tweets and among ridicule tweets is shown in Figure 4. Tweets containing ridicule were less likely to be posted by organizations and were less likely to be informative. Among private users (i.e., Twitter accounts not representing an organization) 51.08% of tweets contained stigma. Across categories, only 1.72% of the classified tweets were classified as representing none of the dimensions, and 69.02% of the tweets were categorized as representing two or more dimensions.

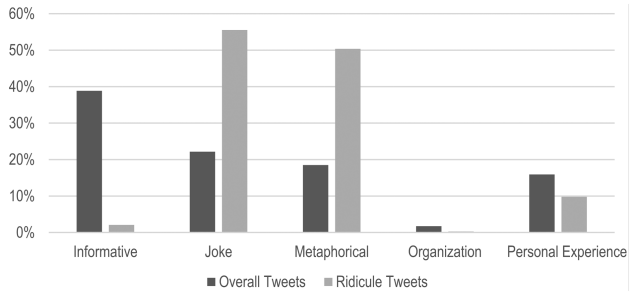

Figure 4. Percentage of machine coded tweets representing each dimension by ridicule. *Note.* Percentages for each dimension are of a total of N = 6,583 ridicule tweets for the light bars and of a total of N = 31,150 overall tweets for the dark bars. 21.13% of all tweets were classified as ridicule by the machine learning model.

Comparison With LIWC Analysis

As expected, the proportion of personal pronouns was related to the tweet being coded by our raters as a personal experience (Table 2). Our dimension "ridicule" was negatively associated with the LIWC measure of positive emotion and also related to the prevalence of personal pronouns (Table 2). Similar results are replicated in the summary correlation table for automated coding results for the entire dataset (Table 3). The exception is an unexpected change in the relationship of "ridicule" to positive and negative emotion, which is further addressed in the discussion.

Discussion

We applied a MLdriven semi-automated coding process to an investigation of AD stigma on Twitter to improve our ability to analyze an otherwise prohibitively large dataset. The conclusion of our methodological focus in this article is that our semiautomated coding procedure replicated manual coding reasonably well even with small input. The substantive finding of our sample analysis of AD stigma was that a substantial proportion of the AD-related posts on Twitter contained ridicule and perpetuated AD stigma, which can impact stereotypes and negative expectations for individuals with the disease and increase "excess disability" (Sabat, 2001). This result contributes to a growing body

Table 2. Correlations and Descriptive Statistics for the Six Stigma Dimensions Coded Manually and Other Dimensions From LIWC Output (N = 311 Tweets)

Variable	1	2	3	4	5	6	7	8
AD stigma								
1. Informative								
2. Joke	−0.46***							
3. Metaphorical	−0.39***	0.21***						
4. Organization	0.43***	−0.27***	−0.21***					
5. Personal	−0.33***	0.20***	0.08	−0.21***				
6. Ridicule	−0.42***	0.47***	0.38***	−0.23***	−0.01			
LIWC								
7. Personal Pronoun	−0.47***	0.47***	0.26***	−0.26***	0.36***	0.35***		
8. Pos. Emo.	−0.10	0.05	0.04	−0.02	0.10	−0.12*	−0.02	
9. Neg. Emo.	−0.16**	0.01	0.03	−0.06	0.07	0.11	0.03	−0.12*

Notes: Alzheimer's disease stigma (AD stigma) variables were 0 = *no*, 1 = *yes* for each tweet. In cases of rater disagreement, a value of 0 was assigned. LIWC variables are the proportion of personal pronouns, positive emotion words (Pos. Emo.), or negative emotion words (Neg. Emo.) per tweet based on the LIWC lexicon (Pennebaker et al., 2007).
*p < .05. **p < .01. ***p < .001.

Table 3. Correlations and Descriptive Statistics for the Six Stigma Dimensions Coded With ML Techniques and Other Dimensions From LIWC Output (N = 31,150 tweets)

Variable	1	2	3	4	5	6	7	8
AD stigma								
1. Informative								
2. Joke	−0.41***							
3. Metaphorical	−0.22***	0.30***						
4. Organization	0.11***	−0.07***	−0.06***					
5. Personal	−0.33***	0.19***	0.00	−0.06***				
6. Ridicule	−0.39***	0.42***	0.43***	−0.06***	−0.09***			
LIWC								
7. Personal Pronoun	−0.46***	0.28***	0.14***	−0.09***	0.19***	0.29***		
8. Pos. Emo.	−0.06***	0.09***	0.07***	0.03***	−0.02**	0.07***	0.05***	
9. Neg. Emo.	−0.27***	0.17***	−0.08***	−0.06***	0.38***	−0.08***	0.08***	−0.14***

Notes: Alzheimer's disease stigma (AD stigma) variables were 0 = *no*, 1 = *yes* for each tweet. LIWC variables are the proportion of personal pronouns, positive emotion words (Pos. Emo.), or negative emotion words (Neg. Emo.) based on the LIWC lexicon (Pennebaker et al., 2007).
*p < .05. **p < .01. ***p < .001.

of work documenting problematic levels of public stigma related to dementia in the United States (e.g., Gove et al, 2016; Riley, Burgener, & Buckwalter, 2014) and internationally (e.g., Batsch & Mittelman, 2012; Blay & Peluso, 2010; Gerritsen, Oyebode, & Gove, 2016). Notably, our investigation makes the novel contribution of utilizing previously unexplored large, publically available social media data. We provide materials and offer recommendations for applying methods similar to those demonstrated here to assist gerontologists in addressing research questions relevant to the social context of aging.

Results showed that the quality of semiautomated coding did vary by dimension. Our algorithm demonstrated superior success for "metaphorical," "informative," and "ridicule" because these three dimensions exhibited high manual rater ICC as well as testing accuracy. Raters were able to reliably parse these dimensions in the text of the tweets, and the automated coding did reasonably well replicating those coding rules when applied to the rest of the tweet dataset. In contrast, "joke" and "personal experience" were more difficult for raters to reliably code (reflected in ICC), which presents limitations for the automated portion of the coding (despite the relatively high testing accuracy on these dimensions). The complex relationship of humor to sentiment was evident in the correlation of joke tweets to both positive and negative emotion. Manual inter-rater reliability for "organization" was the best, however the testing accuracy was poorest on this dimension. Notably, our algorithm's testing accuracy for "organization" did not surpass the accuracy of the SMC by as much as the other five dimensions. The results on this dimension likely stem from the computational difficulty of determining whether a tweet is authored on behalf of an organization or not based on just a single bit of text. For this dimension in particular, contextual information is even more limited because the user's handle may be the only relevant information. Also, raters often relied on prior knowledge unavailable to the ML classifier in order to make a judgment (the meaning of words in the handle, the interpretation of unicode symbols, etc.). A larger training sample is likely the best way for the classifier to improve on this dimension.

For this project, we manually coded a pseudorandomly selected 1% of the total dataset and found, based on the testing accuracy results, that this amount was enough to reasonably approximate the population of tweets for all six dimensions. However, as explained above, less reliable input (i.e., "joke" and "organization") limits the validity of automated coding, even if testing accuracy is high. In general, more training input would improve testing accuracy and would provide a more representative sample which in turn would improve the generalizability of the classifiers. This is illustrated by our observed change in correlation between the LIWC dimension positive emotion, and our dimension, "ridicule." In the manually coded tweets, ridicule was negatively associated with positive emotion (−0.12, p < .05). However, in the entire dataset of tweets with coding predicted by the ML classifiers ridicule was positively associated with positive emotion (albeit very slightly, 0.07, p < .001). Our small sample size may not have been large enough to converge on the proportion of ridicule tweets in the total dataset, and this constituted a limitation in our study. In addition, the relatively weak association of our ridicule dimension with the LIWC negative emotion dimension may be attributed to instances of ridicule that do not necessarily involve emotion words from the LIWC dictionary (e.g., the tweet, "Ric is demented scum").

Our inclusion of LIWC results provided a comparison of our nuanced stigma-related dimensions to typical dimensions in sentiment analysis: positive and negative emotion. These results affirmed the difficulties encountered in defining and coding "joke." This dimension was ambiguously related to positive and negative emotion in the manual and automatically coded datasets. Such results

are consistent with others' attempts to include linguistically complex dimensions such as sarcasm in sentiment analysis (González-Ibánez, 2011). We also used the LIWC dimension of personal pronouns and found that it was positively associated with AD stigma (i.e., ridicule). This appears consistent with past research linking frequency of first person pronouns to depressive symptomology and negative perceptions (as cited in Tausczik & Pennebaker, 2010).

Future Possibilities

There are a number of opportunities for ML technology to further interface with social sciences to yield more accessible tools for researchers. One possibility is to create analysis packages based on well-defined and tested models. For example, our algorithm could be packaged as a content analysis program capable of coding tweets for the six dimensions related to stigma. In this case, the applicability of the program is narrow in terms of specialization to address only these dimensions, and would be validated for use with respect to tweets text samples only. The advantage would be usability, as no manual coding input would be required. Higher user friendliness may entail a tradeoff of lower range of applicability.

A second possibility is that packages could be created that enable researchers more flexibility to train algorithms to code dimensions of their choosing based on their own manual input, as we have done in our demonstration. This information could then be used to train the algorithm capable of coding the remainder of the large dataset in an automated fashion. Built in features could enable researchers to assess the reliability and spot check for validity of that coding. After the initial models are specified, feeding the algorithm additional manual coding input could improve model performance. Although we provide the source code and other materials needed to replicate our methodology, in the future, further packaging of the tool in a simplified interface would increase accessibility.

Recommendations

The software we used to conduct our analysis is included as supplement to this article (see https://github.com/nels-o/ad_stigma for these materials), including commentary on the process and a discussion of the underlying libraries. Several conditions are necessary for replication of the coding method demonstrated here. Related to data collection, depending on the desired dataset size, time on the order of weeks or months may be required to obtain a desired sample. Since 2011 Twitter has increased restrictions on terms of use for data. Keyword searches for tweets are rate limited and it is not generally possible to determine the proportion of the overall set of tweets that are returned.

Also critical are considerations related to the selection, implementation, and evaluation of the ML models. Our initial plan was to use create a simple Naïve Bayes classifier (Zhang, 2004), which has been proven to be a useful tool for text classification problems of all sorts (notably in email spam filtering). After examining this and other options, we used an alternative classification pipeline, which we found to demonstrate superior testing accuracy in our sample study.

Finally, related to analysis of results, cleaning the data is a critical step with large sets of tweets. Large datasets can result in amplified error without attention to removing content that fits the tweet selection criteria but that is unequivocally irrelevant. In our study, tweets related to the release of a popular song titled "senile" accounted for a disproportionate amount of content and would have significantly distorted results if not removed. In another general example, failure to filter out system critical error messages (e.g., those supplied by a phone carrier in a study of cell phone texts) can lead to the false sentiment analysis conclusion that users were angrier than they were in reality.

Conclusion

Although interdisciplinary approaches are recommended for obtaining and analyzing social media data (Bruns & Liang, 2012), it is also true that only basic programming skills are necessary for researchers from a noncomputer science background to employ ML techniques in Python. Many researchers have used custom-made research tools that remain unavailable to others (Bruns & Liang, 2012), and in response to this we have documented our methodology in this article and provide resources enabling replication. The results of our analysis of AD stigma contributes to evidence that, increasingly, social media is a means of expressing attitudes and can impact stigma associated with aging and health conditions (McNeil et al., 2012; Levy, Chung, Bedford, & Navrazhina, 2013). Stigma related to AD and other dementias negatively impacts individuals with dementia, and their friends, family and caretakers through association (Werner & Heinik, 2008). Raised awareness of the power of social media to shape attitudes, and the potential of social media to function as a source of support and information for affected individuals, may eventually result in programs designed for positive impact.

Supplementary Material

Supplementary data is available at *The Journals of Gerontology, Series B: Psychological Sciences and Social Sciences* online.

Funding

This work was supported by the NSF IGERT in Aging Sciences (DGE 0956820).

References

Anderson, N. H. (1968). Likableness ratings of 555 personality-trait words. *Journal of Personality and Social Psychology*, **9**, 272–279. doi:10.1037/h0025907

"*API Overview.*" (2016). Retrieved from https://dev.twitter.com/overview/api

Batsch, N. L., & Mittelman, M. S. (2012). *World Alzheimer Report 2012. Overcoming the Stigma of Dementia*. London: Alzheimer's Disease International.

Bergstra, J., Bardenet, R., Bengio, Y., & Kegl, B. (2011). Algorithms for hyper-parameter optimization. *Advances in Neural Information Processing Systems*, **24**, 2546–2554. doi:2012arXiv1206.2944S

Blay, S. L., & Peluso, É. T. P. (2010). Public stigma: The community's tolerance of Alzheimer disease. *The American Journal of Geriatric Psychiatry*, **18**, 163–171. doi:10.1097/JGP.0b013e3181bea900

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

Breiman, L. (1996). Bagging predictors. *Machine Learning*, **24**, 123–140.

Bruns, A., & Liang, Y. E. (2012). Tools and methods for capturing Twitter data during natural disasters. *First Monday*, **17**. Retrieved from http://journals.uic.edu/ojs/index.php/fm/issue/view/363

Burgener, S., Buckwalter, K., Perkhounkova, Y., & Liu, Y. (2015). The effects of perceived stigma on quality of life outcomes in persons with early-stage dementia: Longitudinal findings part 2. *Dementia*, **14**: 609–632. doi:10.1177/1471301213504202

Cassa, C., Chunara, R., Mandl, K., & Brownstein, J. (2013). Twitter as a sentinel in emergency situations: Lessons from the Boston marathon explosions. *PLoS Currents*, **5**. doi:10.1371/currents.dis.ad70cd1c8bc585e9470046cde334ee4b

Chung, J., & Mustafaraj, E. (2011). Can collective sentiment expressed on twitter predict political elections? *AAAI*, **11**, 1770–1771.

De Choudhury, M., Counts, S., & Horvitz, E. (2013). Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 3267–3276). New York, NY: ACM.

De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. *ICWSM*, **13**, 1–10.

Gerritsen, D., Oyebode, J., & Gove, D. (2016). Ethical implications of the perception and portrayal of dementia. *Dementia*. doi:10.1177/1471301216654036

González-Ibánez, R. (2011). Identifying sarcasm in Twitter: A closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Short Papers* (pp. 581–586). Portland, OR: Association for Computational Linguistics.

Hall, M., Witten, I., & Frank, E. (2011). *Data mining: Practical machine learning tools and techniques* (3rd ed.). Boston, MA: Morgan Kaufman.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, **8**, 23–34.

Isaac, B. M. (2016, February 10). Twitter user growth stalls, and the chief pledges to make fixes. *New York Times*.

Jansen, B., & Zhang, M. (2009). Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, **60**, 2169–2188. doi:10.1002/asi.21149

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web (IW3C2)* (pp. 591–600). New York: ACM.

Levy, B. R., Chung, P. H., Bedford, T., & Navrazhina, K. (2013). Facebook as a site for negative age stereotypes. *The Gerontologist*, **52**, 172–176. doi:10.1093/geront/gns194

McNeil, K., Brna, P. M., & Gordon, K. E. (2012). Epilepsy in the Twitter era: A need to re-tweet the way we think about seizures. *Epilepsy & Behavior*, **23**, 127–130. doi:10.1016/j.yebeh.2011.10.020

Mitchell, T. (1997). *Machine learning*. New York: McGraw Hill.

Mukadam, N., & Livingston, G. (2012). Reducing the stigma associated with dementia: Approaches and goals. *Aging Health*, **8**, 377–386. doi:10.2217/ahe.12.42

Nielsen, F. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *CEUR Workshop Proceedings*, **718**, 93–98.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, **2**, 1–135.

Park, M., Cha, C., & Cha, M. (2012). Depressive moods of users portrayed in Twitter. In *Proceedings of the ACM SIGKDD Workshop on Healthcare Informatics (HI-KDD)* (pp. 1–8). New York, NY: ACM.

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). Linguistic inquiry and word count: LIWC [Computer software]. Austin, TX. Retrieved from www.liwc.net.

Riley, R.J., Burgener, S., & Buckwalter KC. (2014). Anxiety and stigma in dementia: A threat to aging in place. *Nursing Clinics of North America*, **49**, 213–231.

Sabat, S. (2001). *The experience of Alzheimer's disease: Life through a tangled veil*. Oxford, MA: Blackwell Publishing.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the 15th Conference on Computational Linguistics* (Vol. **1**, pp. 172–176). Stroudsburg, PA: Association for Computational Linguistics.

Signorini, A., Segre, A., & Polgreen, P. (2011). The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic. *PLoS One*, **6**. doi:10.1371/journal.pone.0019467

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, **29**, 24–54. doi:10.1177/0261927X09351676

Tumasjan, A., Sprenger, T., Sandner, P., & Welpe, I. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, **10**, 178–185.

Werner, P., & Heinik, J. (2008). Stigma by association and Alzheimer's disease. *Aging and Mental Health*, **12**, 92–99.

Williams, S. A., Terras, M., & Warwick, C. (2013a). How twitter is studied in the medical professions: A classification of Twitter papers indexed in PubMed. *Medicine 2.0*, **2**, e2. doi:10.2196/med20.2269

Williams, S., Terras, M., & Warwick, C. (2013b). What do people study when they study Twitter? Classifying Twitter related

academic papers. *Journal of Documentation*, **69**, 1–74. doi:10.1108/JD-03-2012-0027

Wilson, R., Gosling, S., & Graham, L. (2012). A review of Facebook research in the social sciences. *Perspectives on Psychological Science*, **7**, 203–220. doi:10.1177/1745691612442904

Zhang, H. (2004). The optimality of naive Bayes. In *Proceedings of the International Florida Artificial Intelligence Research Society Conference* (Vol **17**, pp. 562–567). Menlo Park, CA: The AAAI Press.

Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., & Liu, B. (2015). Combining lexicon-based and learning-based methods for Twitter sentiment analysis. *International Journal of Electronics, Communication and Soft Computing Science & Engineering (IJECSCSE)*, **89**, 1–8.

Zimmer, M., & Proferes, N. J. (2014). A topology of Twitter research: Disciplines, methods, and ethics. *Aslib Journal of Information Management*, **66**, 250–261. doi:10.1108/AJIM- 09-2013-0083