Using social media data to predict survey responses: A comparison to multiple imputation

Ashley Richards[1], Joe Murphy[1], Darryl Creel[1]
Justin Landwehr[1]
[1]RTI International, 3040 Cornwallis Road, Research Triangle Park, NC 27709

Abstract

Social media represents a potential auxiliary data source about survey respondents that can be used when measures of interest are not obtained or are missing. Often, these omissions occur from item nonresponse or when survey constraints limit the information that can be collected. In the event that survey respondents share information related to key survey outcomes in their social media postings and allow the researchers access to these data, there is potential to infer the survey outcomes from the social media data. While there are possible pitfalls with such an approach (selection bias, social desirability effects, etc.), social media may serve as a valuable source for these missing data.

To investigate the validity of Twitter data compared to more traditional methods of deriving values for missing data, we compare the results of data predicted from multiple imputation with data predicted from respondents' Twitter posts using two methods: human coders and a machine algorithm. By randomly selecting cases with non-missing values and masking them to the analysis, we are able to use the survey data as a gold standard to evaluate the results of the different methods.

Key Words: social media, imputation, missing data, machine learning

1. Introduction

Missing data is a common problem in survey research. This can be due to sample members who do not participate in the survey at all (unit nonresponse) or participate in only part of it (item nonresponse). While not missing in the traditional sense, data may also be missing because they were not asked for. As nearly anyone who has created a survey has experienced, the researcher often would like to ask more questions than is reasonable in a single survey. Thus, some questions are left out that are still of interest to the researchers.

A common approach for addressing item nonresponse is to derive estimates for the missing values through statistical imputation techniques. While imputation is an accepted approach to adjust for item nonresponse, the approach cannot estimate values for items that are missing across *all* respondents, such as items that were not included in a survey. Occasionally, these data are available on rich sampling frames, but more often,

supplementary data about respondents are difficult if not impossible to acquire while staying within the constraints of a research budget.

As technology becomes more ingrained in our lives, the amount of available data about each of us increases at a seemingly exponential rate. Social networking, in particular, is a source of rich data about its users. Given that 74% of online adults use social networking sites, we can assume many of our survey respondents are sharing information about themselves in this way (Pew Research Center, n.d.). Although, we are aware of the quality concerns inherent in social media data, including undercoverage, differential use, social desirability factors, and more (AAPOR, 2014). Social media represents a potential auxiliary data source about survey respondents that can be used when measures of interest are missing in a traditional or nontraditional way.

Our focus of this investigation was how social media data can supplement survey data when measures of interest are missing. In particular, we wanted discover if missing information could be captured through an analysis of respondents' Tweets. Can respondents' Tweets be used to predict characteristics of individual respondents including sex, age, income, political views, and health status?

We investigated the validity of Twitter data compared to a traditional method of deriving values for missing data. The results of data predicted from multiple imputation were compared with data predicted from respondents' Tweets using two less-traditional methods: human coders and a machine algorithm. Our aim was to see how these approaches compare to imputation since imputation is not always possible, e.g. in the case where questions of interest are not asked. We randomly selected cases with non-missing values and masked them to the analysis in order to use the survey data as a gold standard to evaluate the different methods.

2. Methods

The data were collected by Web as part of an omnibus survey that measured attitudes across a variety of topics, including health, crime, politics, and technology. The survey targeted a general population of U.S. adults age 18+. A total of 2,119 respondents completed the survey out of a sample of 3,623 for a cooperation rate of 58%. The survey was offered in English using sample from KnowledgePanel® and was fielded between March 6 and March 18, 2013.

A total of 12% of the survey respondents said they have used Twitter to post their own Tweets and 19% said they have used it to read Tweets by others. We asked those who had posted their own Tweets if they would provide their Twitter handle (i.e. username) and allow us to merge their public Tweets and survey responses for analysis purposes. A total of 42% (5% of all respondents) said we could. We then used the twitteR package for R (Gentry, 2014) to tap into the Twitter API and collect up to 1,000 of the most recent Tweets from these respondents (or fewer if they did not Tweet this much). The mean number of available Tweets per respondent was 248 and the median was 78.

We randomly divided the respondents with Twitter data into two groups. One group's Tweets and responses were used to train the machine algorithm; the other group's responses were masked and their Tweets were used to predict these artificially missing responses. The six masked items were categorical variables: sex, age, income, health status, depression symptomology, and who they voted for in the last U.S. presidential election.

2.1 Human Coders

Three human coders from our professional research staff predicted values of the six artificially missing survey items at the respondent level. Inspired by the age and weight-estimating carnival game "Fool the Guesser," we set it up like a game to motivate the coders and increase engagement with the task (*Figure 1*). This approach was inspired by gamification research techniques (e.g. Puleston, 2013). We provided the reviewers with the 1,000 most recent Tweets for each respondent (or fewer if the respondent did not have 1,000 total Tweets). They were given five minutes per respondent to skim through the Tweets and make a prediction of the six characteristics of interest using pre-specified ranges for each characteristic. We set the time limit to five minutes to keep the task efficient. More time allotted may have resulted in better predictions.

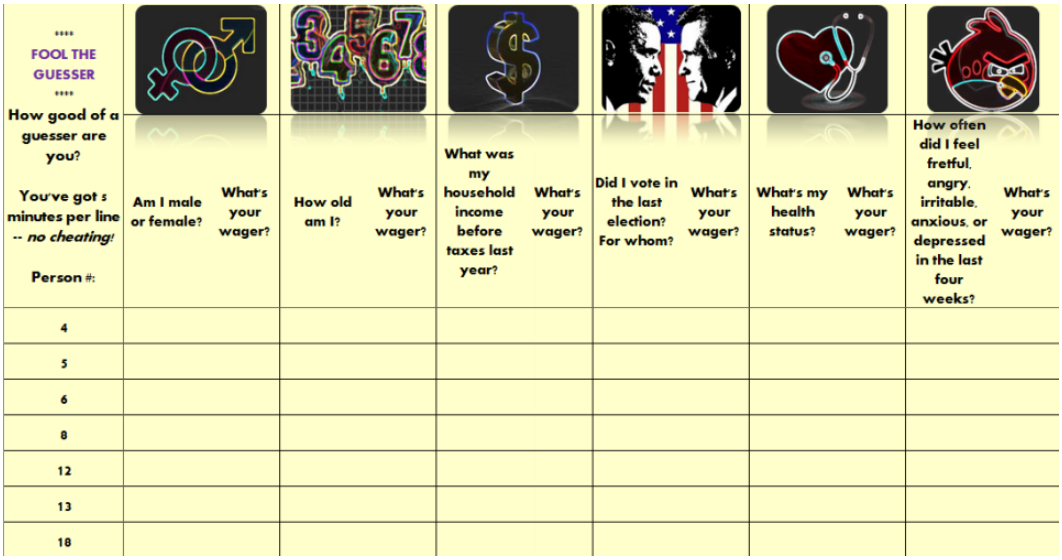

Figure 1: Fool the Guesser prediction sheet

2.2 Machine Algorithm

The second approach we used to predict values of the artificially missing items was a computer prediction. We analyzed the Tweets and survey responses using a computer algorithm based on the k-nearest neighbors method (k-NN), which makes predictions based on multiple data points and cases that resemble other cases.

The first step was cleaning the data. We removed Twitter users from the sample if they had fewer than three Tweets or if their Tweets were not in English. Half of the dataset was used to train the text mining algorithm to predict survey responses based on patterns in Tweet data, and the remaining half was used to test the algorithm's performance.

k-NN is a common and simple supervised learning algorithm. Supervised learning means that the outcome is known ahead of time and the model is built using these example outcomes. k-NN works by projecting points into multi-dimensional space and predicting which category a new point falls into based on the points closest to it (i.e., the "neighbors" that are "nearest" to it). The algorithm does not rely on any statistical models to predict new observations. Predictions are based entirely on patterns in the data (Miner et al., 2012).

2.3 Multiple Imputation

We also predicted values of the artificially missing items through multiple imputation. Five imputations were created using the mice package (Van Buuren & Groothuis-Oudshoorn, 2011) in R (R Core Team, 2013). In mice, the CART method was used to impute. The CART method had an extensive list of variables from which to choose to create the trees.

Multiple imputation is a common approach for predicting missing values, but we used it in an unconventional way. Because the human coding process had multiple coders, we used the multiple imputations to mimic the coding process and produce estimates for missing data at the respondent level.

3. Results

The machine algorithm resulted in a single prediction per variable. The human coders produced three predictions (one per coder) and multiple imputation produced a set of five plausible predictions. We created two new variables for each respondent to indicate what percentage of time the human coders and imputation predicted a value that matched the true value. When comparing the three approaches, we present the mean accuracy, i.e. the percentage of time the predicted value matched the true value. In addition to comparing the three approaches to each other, we also compared them to the accuracy of a purely random prediction to see how much better they performed. The mean accuracy for the random prediction was calculated as $\frac{1}{R}$ where R is the number of response options for a given item. The accuracy of the approaches is presented in *Table 1*.

Table 1: Percentage of correct predictions of missing values by topic and prediction method

	Random	Human	Machine	Imputation
Voting	25.0	58.6	65.5	50.3
Health	20.0	27.6	31.0	29.7
Depression	20.0	26.4	44.8	33.8
Income	25.0	26.4	25.0	31.7
Age	25.0	49.4	31.0	47.6
Sex	50.0	77.0*	48.3*	54.5

*The difference between human coders and the machine algorithm was significant: $p < .05$.

Overall, the human coders and machine algorithm were mostly better than random in predicting the six characteristics. Human coders were the most accurate of the approaches in predicting sex and age. These characteristics seemed rather evident to the coders in reading through the content of the Tweets. The machine algorithm was more accurate in predicting "hidden" characteristics like health and depression, which were infrequently discussed outright in Tweets, but were detectable in patterns from other variables in the algorithm and response data. The machine algorithm was also more accurate in predicting who respondents voted for. Neither approach was very accurate in predicting income; both performed about as well as expected if randomly selecting a value.

We conducted two-proportion z-tests to test for statistically significant differences in the accuracy of each approach compared to the other approaches. The only significant difference was in the human versus machine algorithm predictions of sex: humans were significantly more accurate than the machine algorithm, $Z = 2.26$, $SEM = .13$, $p < .05$ (two-tailed).

One goal of this research was to see how human coders and the machine algorithm compare to the more traditional approach of imputing missing values. The results were mixed for demographic variables: human coders were more accurate in predicting age and sex but imputation was slightly more accurate in predicting income. The machine algorithm performed worse than imputation on the demographic items. Most of the differences were not statistically significant.

For the three other variables – voting, health, and depression – the machine algorithm was the most accurate approach and consistently outperformed imputation. Human coders were more accurate than imputation in predicting who respondents voted for and imputation was better at predicting health and depression, but the differences were not statistically significant.

4. Discussion

We found that even with a small set of respondents, Tweets can be used to gather additional information about respondents and predict missing values. For age, sex, and income, predictions by human coders were more accurate than predictions by the machine algorithm, but the differences were only statistically significant for sex. For more "hidden" characteristics, the machine algorithm outperformed human coders but not by statistically significant margins. We compared the accuracy of these approaches to imputation, which is a more traditional and commonly used approach. The findings were mixed. Neither approach consistently outperformed imputation, but imputation was not consistently most accurate. Once again, the differences were not statistically significant.

Our findings are limited by the small sample size. We had Tweets for only 5% of all respondents. Tweets from half of those respondents were used to train the machine algorithm, leaving us with only 29 respondents to make predictions about. A larger sample would have allowed us to better test for differences between the methods.

Both of the non-traditional approaches could be improved to potentially produce more accurate results. The human coder approach could be improved by selecting coders with a keen sense of perception and an eye for detail. Additionally, their performance might improve through training, experience, and if given more time to review the Tweets for each case. As for the machine algorithm, this approach may result in greater predictive accuracy if a larger sample of respondents with Tweets were provided so the algorithm could better learn from them.

Although these approaches may be useful in the future, they are not without limitation. Relying on Twitter data is challenging because many respondents are not on Twitter. While Twitter and other social media are gaining in popularity, not everyone uses them and those who do are more likely to be younger, female, and better educated (Duggan & Smith, 2013).

Even among respondents who are on Twitter, we found that many did not give permission for us to use their Tweets. In some cases it may be possible to find publicly-

available Tweets from respondents even when they do not provide their Twitter handle and give permission for their Tweets to be used, but that raises an ethical issue and we thought it would be best to only use Tweets with permission.

We were encouraged by the fact that results from Tweets, either predicted by humans or a machine algorithm, could produce estimates with accuracy in the same range as imputation procedures. This is important because imputation cannot be used in the case where survey questions were not asked. For this reason, and despite the known limitations, Twitter may represent an important resource moving forward in the estimation of values desired in a survey, but not asked because of space limitations, the desire to reduce respondent burden, or other factors.

References

American Association for Public Opinion Research (AAPOR), Task Force on Emerging Technologies in Public Opinion Research. (2014). *Social Media in Public Opinion Research: Report of the AAPOR Task Force on Emerging Technologies in Public Opinion Research.* Retrieved from http://www.aapor.org/Social_Media_Task_Force_Report.htm.

Duggan, M. & Smith, A. (2013). Social Media Update 2013. Pew Research Center. http://www.pewinternet.org/files/2013/12/PIP_Social-Networking-2013.pdf

Miner, G., Delen, D., Elder, J., Fast, A., & Nisbet, R. A. (2012). *Practical text mining and statistical analysis for non-structured text data applications*. Waltham, MA: Academic Press.

Gentry, J. (2014). *Package 'twitteR.'* Retrieved July 23, 2014 from http://cran.r-project.org/web/packages/twitteR/twitteR.pdf

Pew Research Center. (n.d.). "Social Networking Fact Sheet." Retrieved July 23, 2014 from http://www.pewinternet.org/fact-sheets/social-networking-fact-sheet/

Puleston, J. (2013). Gamification of Market Research. In C. A. Hill, E. Dean, and J. Murphy (Eds.), *Social Media, Sociality, and Survey Research* (pp. 253-294). Hoboken, NJ: John Wiley & Sons.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

Van Buuren, S. & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equestions in R. *Journal of Statistical Software*, 45*(3)*, 1-67. http://www.jstatsoft.org/v45/i03/