Roger Qiyuan Jin

POLS 8500

Spring 2017

# Measuring Government Performance Using Sentiments of Tweets: Whether Big Data Match Citizen and Employee Surveys

## Introduction

The difficulties of measuring government performance have been well documented in public management literature (Boyne and Gould-Williams 2003; Moynihan 2008). Assessing performance in the public sector remains a key question to public management, in which an entire subfield of performance measurement is dedicated to tackle the problem (Meier and O'Toole 2013).Majority of previous studies that focused the performance of federal government have relied on employees' or managers' self-assessment of performance of their agencies in surveys like Federal Employee Viewpoint Survey, or Merit Principles Survey. Studies using "objective" measures of government performance often adopt proxy measures or indicators (test scores, crime rates, etc.) that only capture parts of what public organizations are doing. Moreover, the cost of collecting these subjective or objective performance data is usually rather high. In this project, I explore the possibility of using "big data" from social media to measure government performance. Specifically, I examine whether sentiments of tweets and general sentiments of tweets about a particular federal agency would match 1) self-reported performance data from the agency employees and 2) citizens' favorability rating of this agency. Tweets mentioning 17 department or independent agencies of federal government are collected and sentiment analyzed using a machine learning algorithm called Support Vector Machine. Sentiment score of each agency are compared with perceptual performance data from 2015 Federal Employee View Point Survey and favorability data from a 2015 political survey conducted by Pew Research Center. However, results show negative correlation between ranking from sentiment scores of tweets and ranking of citizens' favorability, and negative correlation between ranking from sentiment scores of tweets and ranking from employees' self-reported performance. The associations are not significant. The null results still provide some important insights to this emerging research area in public administration. I will first discuss the data and method used in the paper, followed with a discussion of the results and conclude with summary of the study and directions for future research.

## Data and Method

### 1. *September 2015 Political Survey*

This paper draws data on bureaucratic reputation from three different sources to answer the research question.

The first source of data is Pew Research Center's September 2015 Political Survey, which asked a national sample of 1,502 adults living in the United States a series of politically-related questions, ranging from job approval ratings of political institutions like U.S. Congress and Supreme Court, to favorability of presidential candidates and federal agencies. The survey was conducted during the time period of Sep 22 – 27, 2015. Statistical results are weighted to correct known demographic discrepancies. The margin of sampling error for the complete set of weighted data is ±2.9 percentage points. The data was downloaded from Pew Research Center's website.

The main variable of interest from the 2015 Political Survey data is respondents' favorability of 17 federal agencies. Each respondent was asked to answer the survey question "Is your overall opinion of [agency name] very favorable, mostly favorable, mostly unfavorable, or very unfavorable?" From this questions, the survey obtained rating of 17 federal agencies including Department of Health and Human Services (HHS), Federal Bureau of Investigation

(FBI), Environmental Protection Agency (EPA), The Postal Service (USPS), Social Security Administration (SSA), Department of Veterans Affairs (VA), Department of Education, National Security Agency(NSA), Department of Homeland Security, Food and Drug Administration (FDA), The Defense Department, Central Intelligence Agency (CIA), Internal Revenue Service (IRS), Centers for Disease Control and Prevention (CDC), The Justice Department, National Aeronautics and Space Administration (NASA), The National Park Service. Using the percentage of respondents who have favorable views towards them, these agencies are ranked from viewed most favorably to most unfavorably (See Table 1). In this survey, the USPS is the most popular agency, with 84% of respondents having favorable views impressions, while the Department of Veteran Affair is the least favored agency, with merely 39% of favorable views. According to a Pew Research Center report, Currently, VA's favorability dropped 29 percentage points since October 2013, during the partial government shutdown. This dire impression is reportedly related to problems with its health care services for veterans, and the resignation of scandal-embattled VA secretary Gen. Eric Shinseki in 2014.

### 2. *Federal Employee Viewpoint Survey 2015*

Ratings of federal agencies by their own employees come from the 2015 Federal Employee Viewpoint Survey, which is administered every other year by the Office of Personnel Management (OPM). This is the latest FEVS survey results, which is released in late 2015. The survey is designed to assess the attitudes of full-time, permanent federal employees with respect to workplace issues. About 84 questions were asked, including opinions of work experience, performance culture, leadership, job satisfaction and demographic information. In 2015, a stratified, representative sample of approximately 848,237 was selected and 421,748 federal employees from 82 agencies participated in the survey, with a response rate of 49.7 percent. For most of the questions, respondents answered on a Likert-scale from 1 to 5, with 1 representing strong disagreement and 5 representing strong agreement. To gauge employees' opinion on whether their agency is a good place to work and whether they are satisfied with their job, this paper focuses on three survey questions: 1) I recommend my organization as a good place to work; 2) Considering everything, how satisfied are you with your job? 3) Considering everything, how satisfied are you with your organization? A scale measuring the employee's satisfaction of their agency is constructed using the responses to these three questions. To match Pew Research Center's survey data, satisfaction data for previously mentioned 17 agencies are aggregated at agency level (using mean of each agency). Unfortunately, Federal Employee Viewpoint Survey does not include employees from USPS so there is no data available for Postal Service in this survey. Based on the aggregated satisfaction score of each agency, a new ranking of these 16 (without USPS) has been produced (See Table 1). In this ranking, NASA was rated the best agency to work in federal government, while Department of Homeland Security has the lowest satisfaction score. Reports from a nonprofit organization called Partnership for Public Service, which produces rankings of best place to work in federal government using similar methodology, have also shown consistently for several years that NASA and DHS at opposite position of the ranking.

### 3. *Twitter Data*

Finally, to measure citizens' view towards federal agencies, tweets mentioning the 17 agencies examined by Pew Research Center (eg. @EPA) are collected from Twitter Advanced Search using a Python web scraping script. Specifically, to match the time frame of the survey, all tweets posted during Sep 22-27, 2015 and also mentioning the 17 agencies in the 2015 Political Survey were collected. Although Twitter API allows developers to easily collect tweets using different search queries, the rate limit of the API would only allow tweets posted within roughly recent two weeks to be collected. To circumvent the rate limit, the author adopted a web scraping script [1] in Python to extract tweets from the Twitter Advanced Search, where Twitter allows users to build a detailed search query, including a time frame, to view certain tweets. For the 17 agencies, 18 search queries (IRS has two Twitter accounts) were built to obtain tweets posted during Sep 22 to Sep 27 and mentioning each agency. After deleting tweets from agency's own account, a total of 21825 tweets were collected. For each tweet, the web scraper was able to extract information such as user ID, date and time of tweet, retweet number, favorites, and most importantly, text of the tweet. Among the agencies, NASA has the most number of tweets (5978) in the time period, and SSA has fewest number of tweets – only 120. Then sentiment analysis was conducted using a supervised machining learning method.

### 4. *Sentiment Score of the Tweets Using Support Vector Machine*

Computing sentiment score or classifying tweets sentiment is fairly common among research on social media data. Although it is often considered a difficult task to analyze the sentiment a large text document, classifying tweets into different categories based on sentiment is often possible given the short length of text (140-character limit). This paper used a supervised learning method called Support Vector Machine to classify and compute sentiment score of the tweets.

This approach consists of the following steps: (1) obtain a training dataset that contains 5513 manually classified tweets, (2) divide the dataset into a training dataset and a test dataset, (3) train a classifier on the training set using Support Vector Machine, (4) apply the SVM

[1] Credits should be attributed to Simon Lindgren, Professor of Sociology at Umeå University in Sweden, who shared his Python code on GitHub (https://github.com/simonlindgren) and kindly answered my questions through email.

classifier on the test set for cross-validation, and check the classifier’s accuracy, specificity and sensitivity, (5) if the classifier works well on the test set, apply the trained classifier on the collected tweets that talk about federal agencies, then obtain a sentiment classification and corresponding probability for each tweet.

The training dataset is called Sanders-Twitter Sentiment Corpus, which is downloaded online at Sanders Analytics. The dataset consists of 5513 hand-classified tweets, which were classified into four different categories – positive, negative, neutral and irrelevant. The content of these tweets are mainly about topics regarding Apple, Google, Microsoft and Twitter. According to the manual of this corpus, all classification was done by an American male who is fluent in English. To increase the accuracy and shorten processing time, the analysis of this paper uses only the classified positive and negative tweets. The trimmed dataset contains 1091 classified tweets, with 572 negative and 519 positive.

The trimmed Sanders Corpus was divided into a training set and a test set, with 75% of the tweets randomly assigned in the training set and 25% in the test set. The texts of the tweets were preprocessed in R with tokenization, formatting, removing stop words and stemming before training the classifier. First, the tweets in the Sander Corpus were converted into a corpus format using *tm* package in R, which effectively tokenizes each of the document (tweet in this case). Then each document was formatted with removing of punctuation, numbers and converting to lower case. URLs in tweets were also removed because they normally do not represent relevant content but rather point to it. All the texts were converted to lower case and punctuation and numbers. Stop words was then removed from each tweet using the *tm* package default English stop words. Lastly, the corpus was processed with a stemming procedure, which reduces every word to its stem. The preprocessed corpus was then converted to a document-term matrix, which is ready to train the classifier.

I trained a sentiment classifier on the training set using Support Vector Machine (SVM), which is a widely used, state-of-the-art supervised learning algorithm, well suited for large scale text categorization tasks, and robust on large feature spaces (Ranco et al. 2015) . As an old but effective machine learning methods, SVM was introduced in 1992 by Vapnik and it usually has great performance in analyses that deal with image recognition, text data, etc. The method fits a maximally separating hyperplane between a set of data points to classify them into different groups. Using package *e1071* in R, a SVM classifier was trained on the training data to distinguish between positive and negative tweets.  A linear kernel was used to in the algorithm, which is well-suited for classifying texts and was tested to produce the best classifier on the training data. Then the classifier was cross-validated using the test dataset. I would describe specifically in the next section, the classifier performed well on the test set, with high accuracy, specificity and sensitivity across the board. Then this cross-validated classifier was applied to the collected Twitter data, which contains 21825 tweets about 17 federal agencies. The collected tweets are preprocessed using a similar procedure described above. For each tweet, the classifier would predict three outcomes: probability of this tweet being positive, probability of this tweet being negative, and a sentiment classification (negative or positive) based the probabilities. For instance, if the probability of a particular tweet being positive is bigger than 0.5, this tweet will be classified as positive. The probability of tweet being positive will be used as sentiment score for the next part of the analysis.

### 5. *Correlation of Three Ranks Using Kendall's tau and Spearman's rho*

After obtaining the sentiment score (probability of being positive) for each tweet, the results are aggregated at the agency level using mean scores. In this way, we can get a measure of how Twitter users view different federal agencies in terms of favorability. Then a third ranking of agencies by favorability/reputation was calculated using the average sentiment score of each agency. At this point, I have obtained ranking of 17 federal agencies using three different data source: Survey from Pew Research Center, Survey of Federal Employees, and sentiments of Twitter users. I am interested in whether these three rankings could be related to each other. To accomplish this goal, I used Kendall's tau and Spearman's rho, which are both considered to be appropriate to assess statistical associations based on the ranks of the data, to measure the correlation of these three rankings. I computed Kendall's tau and Spearman's rho for three pairs of the rankings and the results will be reported in the next section.

**Table 1. Summary Statistics for Federal Agencies**

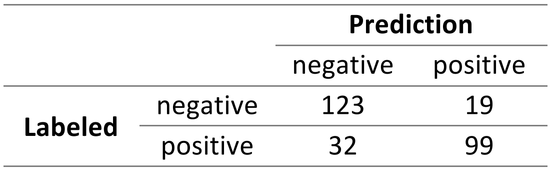
| Agency | positive | negative | Rank Sentiments | Rank Pew | FEVS | Rank Fevs | N |
|---|---|---|---|---|---|---|---|
| *usps* | 0.544 | 0.456 | 16 | 1 | NA | NA | 1517 |
| *natl park* | 0.587 | 0.413 | 13 | 2 | 53.1 | 15 | 415 |
| *cdc* | 0.604 | 0.396 | 3 | 3 | 70.9 | 2 | 1085 |
| *nasa* | 0.598 | 0.402 | 6 | 4 | 76.1 | 1 | 5978 |
| *fbi* | 0.588 | 0.412 | 12 | 5 | 69.9 | 3 | 3437 |
| *dhs* | 0.590 | 0.410 | 10 | 6 | 43.1 | 16 | 1276 |
| *dod* | 0.585 | 0.415 | 14 | 7 | 58.4 | 12 | 839 |
| *cia* | 0.597 | 0.403 | 7 | 8 | 67.1 | 4 | 1142 |
| *ssa* | 0.590 | 0.410 | 9 | 9 | 66 | 8 | 120 |
| *hhs* | 0.641 | 0.359 | 2 | 10 | 63.9 | 9 | 306 |
| *nsa* | 0.580 | 0.420 | 15 | 11 | 67 | 5 | 374 |
| *epa* | 0.601 | 0.399 | 5 | 12 | 58.5 | 11 | 1721 |
| *fda* | 0.526 | 0.474 | 17 | 13 | 66.3 | 6 | 734 |
| *doj* | 0.590 | 0.410 | 11 | 14 | 66.2 | 7 | 1035 |
| *edu* | 0.592 | 0.408 | 8 | 15 | 61.3 | 10 | 768 |
| *irs* | 0.690 | 0.310 | 1 | 16 | 55.5 | 13 | 518 |
| *va* | 0.603 | 0.397 | 4 | 17 | 55.1 | 14 | 560 |
| *All* | 0.592 | 0.408 |  |  |  |  | 21825 |

## Results

### *1. SVM Classifier*

I will first report the performance of the SVM classifier trained using the Sanders Twitter Corpus. After applied the trained classifier to the test dataset, I compared the results with the hand-classified results, which was shown in the following confusion matrix (See Table 2). It is easily to tell that the classifier performed well on the test set, with most of the negative tweets classified as negative and most of the positive tweets classified as positive. I also computed the accuracy, specificity and sensitivity scores for the classifier, which all indicate good performance. The accuracy of the classifier's prediction on the test set is 0.81, which means that overall 81% of the tweets were classified correctly. The specificity of the classifier's prediction is 0.87, meaning 87% of the negative tweets are correctly classified. Moreover, the sensitivity of the prediction is 75%, indicating that 75% percent of the positive tweets are correctly identified. Overall, the results from the confusion matrix and scores of accuracy, specificity and sensitivity all show that the trained SVM classifier has great predicting power.

**Table 2. Confusion Matrix for SVM prediction**



| | | **Prediction** | |
|---|---|---|---|
| | | negative | positive |
| **Labeled** | negative | 123 | 19 |
| | positive | 32 | 99 |

### *2. Classifying Collected Twitter Data and Report Sentiment Scores*

After training the SVM classifier and preprocessing the collected tweets about federal agencies, the classifier was then applied to the 21825 tweets collected. For each tweet, the classifier would produce three outcomes: probability of tweet being positive, probability of tweet being negative and sentiment classification based the probabilities. In general, the results show 5672 tweets were classified as negative, 16153 tweets were classified as positive. The average probability of a tweet being positive is 0.592, and the average probability of a tweet being negative is 0.408. By observing the probability of tweets being positive, we know the agency that is viewed most favorably is IRS, with an average probability of 0.69 of mentioned tweets being positive. The least favored agency by Twitter users is USPS, with an average probability of 0.54 of being discussed favorably. Also using the probability of tweets being positive as sentiment score, I ranked the agency from viewed most positively to most negatively (See Table 1).

### *3. Correlation of Three Ranks Using Kendall's tau and Spearman's rho*

After obtaining a ranking that used sentiment score of tweets mentioned each agency, I now have three rankings based on data from citizen survey, employee survey and Twitter. A plot of the three rankings is shown in Figure 1. To examine whether these ranking are potentially related, I use Kendall's tau and Spearman's rho to test the statistically associations between them.

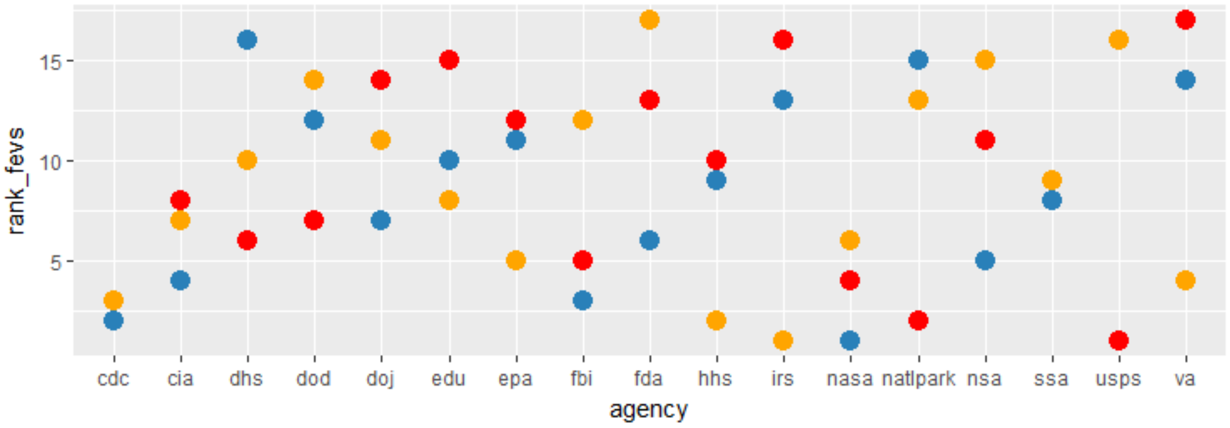


**Figure 1. Plot of Agency Rankings**

The results from the correlation test shows (see Table 3) that the ranking from the Twitter data has a negative association with ranking from FEVS and ranking from Pew Research Center's Survey data. And the negative correlations are not statistically significant. The ranking from FEVS and ranking from Pew Research Center have a positive association, but the association is also not statistically significant. The results here show that the citizens' opinion gathered from Twitter data does not match opinions from a national survey or employee's self-assessment of their agency.

**Table 3. Results of Correlation Test**

| | Kendall's Tau | Spearman's Rho |
|---|---:|---:|
| **FEVS v. Twitter** | -0.05 (.82) | -0.04 (.87) |
| **Pew v. Twitter** | -0.22 (0.24) | -0.30 (0.24) |
| **FEVS v. Pew** | 0.28 (.14) | 0.28 (.29) |

## Conclusion

This paper is an exploratory study that tries to use social media data to measure government performance. Tweets about 17 federal agencies are collected and sentiment analyzed to match citizen survey and employee survey. By comparing the ranking of the agency from three different data source, no positive association was found between the Twitter sentiments and survey data. The null result does not necessarily mean Twitter data are not valuable for performance related research. The result is probably caused by the specific research design of this paper. The limited time period of the Twitter data collected might contribute to the representativeness of the data. Also, the result may indeed suggest a different user base of Twitter users, which probably hold different opinions towards federal government (Mergel 2013a, 2013b). Another possible explanation for the discrepancies might be people who choose to tweet about certain federal agency might have positive or negative bias toward the agency. For example, large amount of tweets that mentioned USPS are complaints about their package delivery service, usually about package loss or delivery delay. This kind of bias of Twitter users and their tendency to complain about service might be helpful for federal agencies as they could use tweets to potentially evaluating performance of certain local units. For example, more complaints about a certain office might indicate poor performance.

This paper is a first step using machine learning techniques to analyze social media data about U.S. federal agencies. It contributes to the field of public management by offering an alternative way of measuring performance of public organizations – using sentiments of performance-related tweets, which is easily accessible, in real-time and less costly. There is great future research potential in this topic area. Future study could collect specific performance-related data to more accurately gauge performance. It could also be useful to explore whether Twitter sentiments could be used as a measure of bureaucratic reputation, which is also an important research topic in public administration and political science. Moreover, it will be quite interesting to explore questions like what federal agencies are tweeting about, whether different type (regulatory, service delivery) of federal agencies tweets about different information, and interaction pattern with citizens (Knox 2016; Mergel 2010; Waters and Williams 2011). As field of public affairs enters era of "big data", this paper answers the call of taking advantage of newly emerging data and meaningfully combining them with administratively collected data to have value in improving public programs (Mergel, Rethemeyer, and Isett 2016) and furthers our knowledge of whether adopting social media increases citizen participation, collaboration and transparency.

**Reference**

Boyne, George, and Julian Gould-Williams. 2003. “Planning and Performance in Public Organizations An Empirical Analysis.” *Public Management Review* 5(1): 115–32.

Knox, Claire Connolly. 2016. “Public Administrators’ Use of Social Media Platforms: Overcoming the Legitimacy Dilemma?” *Administration & Society* 48(4): 477–96.

Meier, Kenneth J., and Laurence J. O’Toole. 2013. “Subjective Organizational Performance and Measurement Error: Common Source Bias and Spurious Relationships.” *Journal of Public Administration Research and Theory* 23(2): 429–56.

Mergel, Ines. 2010. “Government 2.0 Revisited: Social Media Strategies in the Public Sector.” *Public Administration*. http://surface.syr.edu/ppa/2.

———. 2013a. “A Framework for Interpreting Social Media Interactions in the Public Sector.” *Government Information Quarterly* 30(4): 327–34.

———. 2013b. “Social Media Adoption and Resulting Tactics in the U.S. Federal Government.” *Government Information Quarterly* 30(2): 123–30.

Mergel, Ines, R. Karl Rethemeyer, and Kimberley Isett. 2016. “Big Data in Public Affairs.” *Public Administration Review* 76(6): 928–37.

Moynihan, Donald P. 2008. *The Dynamics of Performance Management: Constructing Information and Reform*. Georgetown University Press.

Ranco, Gabriele et al. 2015. “The Effects of Twitter Sentiment on Stock Price Returns.” *PLOS ONE* 10(9): e0138441.

Waters, Richard D., and Jensen M. Williams. 2011. “Squawking, Tweeting, Cooing, and Hooting: Analyzing the Communication Patterns of Government Agencies on Twitter.” *Journal of Public Affairs* 11(4): 353–63.