Articles

Text Mining in Survey Data

Christine P. Chai*

Keywords: topic modeling, free text responses, text mining, surveys, supervised latent dirichlet allocation, open-ended questions

DOI: 10.29115/SP-2018-0035

Survey Practice

Vol. 12, Issue 1, 2019

Free text responses in surveys contain important information and should be analyzed by researchers. However, human coding of survey text is not only expensive, but also vulnerable to subjectivity. An automated text mining approach can solve these problems. Therefore, we demonstrate using the supervised latent Dirichlet allocation (sLDA) to jointly analyze text and numerical data in an employee satisfaction survey. For each rating, the algorithm outputs selected words as the "topic" and estimates the credible interval. Finally, we discuss future applications and advantages of utilizing survey text.

DISCLAIMER

The research was conducted while the author was a PhD student in statistical science at Duke University. The views and opinions expressed in this manuscript are those of the author and do not reflect the view of the U.S. Department of Housing and Urban Development or of the U.S. government.

INTRODUCTION

Most surveys contain both text and numerical data, but people often follow certain survey data analysis guidelines (SurveyMonkey, n.d.; Statistical Services Centre 2001) and focus only on the numbers. This is understandable because text data are unstructured, and text is generally more difficult to analyze than numerical answers (Schuman and Presser 1996).

However, free text responses can provide more diverse explanations of respondents' experience than the numerical counterparts. For example, free text responses can serve as alternative explanations to "why this option was selected" (Jackson and Trochim 2002). For another example, open-ended questions are better at capturing the motivations of family forest owners than the fixed-response questions (Bengston, Asah, and Butler 2011).

Nowadays, survey text analysis is still relatively rare, and when conducted, it is often done manually (Roberts et al. 2014), which tends to be expensive (Grimmer and Stewart 2013). Moreover, human coding in surveys is subjective and prone to intracoder variability, even in trained, experienced professionals

* **Institution:** U.S. Department of Housing and Urban Development **Department:** Office of Risk Management

(Patel et al. 2012; Yamanishi and Li 2002). On the contrary, an automated text coding system does not suffer from the inconsistencies that human coders do, so incorporating text mining in survey analysis would be useful for extracting information from the free text responses.

Another advantage of text mining is to pick up trends rarely noticed by humans. For example, (Schuman 2008) focused on how responders used pronouns as a way to predict how they felt about the U.S. government during the Vietnam War. Those who referred to the United States as "we" were more supportive of the U.S. involvement, while the respondents who used "they" were more likely to be against the U.S. government's participation.

Given these benefits, text mining has been applied successfully in many settings. Many text mining algorithms are unsupervised, including latent Dirichlet allocation (Blei, Ng, and Jordan 2003) and pattern clustering (Quan, Wang, and Ren 2014). These algorithms use text as the sole input and identify topics from the corpus.

Nevertheless, when numerical ratings are present in the text corpus, they should also be utilized. The sLDA (supervised latent Dirichlet allocation) is a solution to combined analysis of text and numerical data; this algorithm uncovers latent topics from a corpus with "labeled" text documents, i.e. each document in the corpus is associated with a rating or a category (Blei and McAuliffe 2007). The sLDA has many existing applications, but most are in the computer science field, such as video activity recognition (Hughes 2010) and credit attribution of bookmarking websites (Ramage et al. 2009).

This article applies sLDA on surveys to jointly analyze text and numerical ratings. We walk through an example of an employee satisfaction survey, provided by Nick Fisher from ValueMetrics[1]. Using the sLDA algorithm from the **R** package *lda* (Chang 2015), we identify ten topics from the text that correspond to the ratings 1-10.

DATASET DESCRIPTION AND PREPARATION

The employee satisfaction dataset contains 530 employee ratings of the company overall on work itself, and a text comment to specify the main reason of their ratings. The data contain 12,475 words, and each comment has 23.54 words on average. The ratings are from 1-10, with 1 the least satisfied and 10 the most. According to the histogram in Figure 1, most ratings are between 5 and 9.

An example of a response with a high score is "The opportunities made available to me in the last year have been enormous. My strengths have been identified and developed." On the contrary, an example associated with a low

1 http://www.valuemetrics.com.au/

score is: "Minimum support in the work. Poor income."

The data contain at least one error — one respondent rated his or her company 1 but with the comment "Love my work — very varied." Clearly the rating scale confused that person, but according to the data provider (Fisher 2013), errors are rare in surveys on people's values (compared with surveys on community values). We also manually checked the comments for ratings 1 and 2; the only obvious error was the record previously mentioned, so we do not think there is a systematic error biasing this dataset.

It is implicitly assumed that everyone rated on the same scale. In reality, the same level of satisfaction can result in different ratings in different people. For example, one may not rate any experience as 10 because he or she thinks no job is perfect, but some do give ratings of 10.

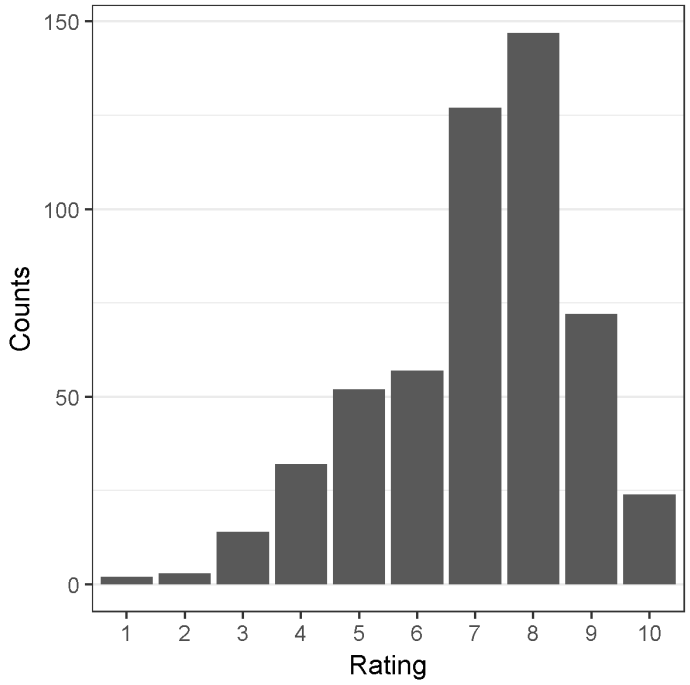

Figure 1: Number of comments in each rating.

DATA CLEANING

To prepare the dataset for analysis, we first needed to reduce the vocabulary size. We achieved this by stemming and tokenizing the words using the *wordStem* function in the **R** package *SnowballC*. This function uses the Porter algorithm (Porter 2001) to assign words of the same stem to the same token. For example, "carry" and its past tense "carried" are assigned to the same token "carri." The Porter algorithm addresses details of English grammar — "fitted" becomes "fit," where the double "t" is removed. It also includes a dictionary to avoid over-stemming — "reply" becomes "repli," not "rep" (abbreviation for "representative"). In this article, the terms "word" and "token" are used interchangeably.

The next step was to remove stopwords (words with little semantic meaning, e.g. "a" or "and"). To simplify the analysis, we also removed punctuation, as recommended by (Francis and Flynn 2010). While certain punctuation, such as repeated exclamation marks, !!!, can be used as an intensifier (Liu 2015), we did not need to address this issue because the data contain only two exclamation marks, both in single form.

ANTONYM REPLACEMENT

Approximately 150 different phrases in the dataset start with a negation term[2], so the order of negation should be retained, since a "not" preceding a word can reverse its meaning (Soriano, Au, and Banks 2013) — e.g. "not bad" means "good," and vice versa. However, most topic models for text data are "bag-of-words" models (Wallach 2006; Zhang, Jin, and Zhou 2010) and do not distinguish word order.

Our solution was to manually identify corresponding antonyms for these phrases and then put the antonyms back into the corpus using the **R** pattern matching and replacement function *grep*. We considered all phrases with the same word but a different negation term to be equivalent — e.g. "not clear" and "isn't clear" are both replaced with "unclear."

TOPIC MODEL: SUPERVISED LATENT DIRICHLET ALLOCATION (SLDA)

To jointly analyze the text and ratings, we implemented the sLDA, a Bayesian data generative process which assigns a topic assignment vector to each word (Blei and McAuliffe 2007). For instance, given three topics, a word's topic assignment vector can be (0.2, 0.5, 0.3). This means the word has proportions 20% in Topic 1, 50% in Topic 2, and 30% in Topic 3. Proportions in topics are defined using word counts.

Implementing sLDA allows us to utilize the Bayesian framework. First, the Bayesian topic model produces credible intervals for each topic, so researchers can directly say "this topic has 68% posterior probability to be in this range of scores." Moreover, in sLDA, each topic is a probabilistic distribution over the words. It is possible to allow certain words (e.g. "supportive") a higher probability in ratings 6-10, so the topic model can "grow" in a particular direction.

ALGORITHM DESCRIPTION

The sLDA algorithm requires a preset number of topics, and it first draws topics from a Dirichlet distribution as the prior, then updates the probabilities using the words in the documents as the likelihood. Finally, sLDA draws the numerical response variable for each document from a normal distribution

2 Negations terms also include the equivalents of "not," such as "no," "isn't," "aren't," "wasn't," "weren't," "don't," "doesn't," and "didn't."

using the posterior topic assignments.

The entire dataset is a set of M documents $D = \{D_1, \cdots, D_M\}$. The words within a document D_d are $W_d = \{W_{d,1}, \cdots, W_{d,N_d}\}$. This means the document D_d contains N_d words, and the total number of words in the dataset is $N = \sum_{d=1}^{M} N_d$. Assume the predefined K topics are represented as a set of vectors $\beta_{1:K} = \{\beta_1, \cdots, \beta_K\}$, and η, σ^2 are preset constants for the normal distributions.

The sLDA process is illustrated in Figure 2 and described below:

For each document D_d,

- Draw topic proportions $\theta_d|\alpha \sim \text{Dirichlet}(\alpha)$
- For each word $W_{d,\,n}$
    - Draw topic assignment $Z_{d,n}|\theta_d \sim \text{Multinomial}(\theta_d)$
    - Draw word $W_{d,n}|Z_{d,n}, \beta_{1:K} \sim \text{Multinomial}(\beta_{Z_{d,n}})$
- Draw response variable $Y_d|Z_{d,\,1:N_d}, \eta, \sigma^2 \sim N(\eta\overline{Z}_d, \sigma^2)$, with $\overline{Z}_d = (1/N_d)\sum_{n=1}^{N_d} Z_{d,n}$

Transform the Gaussian response variables $Y_1, \cdots, Y_M \in R$ to the K categories (preset topics)

- $-\infty = \tau_0 < \tau_1 < \tau_2 < \cdots < \tau_{K-1} < \tau_K = \infty$
- In the k th category, $\tau_{k-1} \leq Y_d \leq \tau_k$

For each document D_d, the initial topic proportions θ_d are determined by the Dirichlet distribution with the preset constant α. For each word $W_{d,\,n}$ in the document D_d, we draw a topic assignment vector $Z_{d,n}$ from the multinomial distribution and the initial topic proportions θ_d. Then the word $W_{d,n}$ is reassigned from another multinomial distribution, given the topic assignment vector $Z_{d,n}$ and the predefined topic vectors $\beta_{1:K}$. After the whole document D_d is processed, the response variable Y_d is drawn from a normal distribution, using preset constants η, σ^2 and the mean of $Z_{d,n}$. Finally, after getting all Gaussian response variables $Y_1, \cdots, Y_M$, sLDA maps them to the K categories (preset topics).

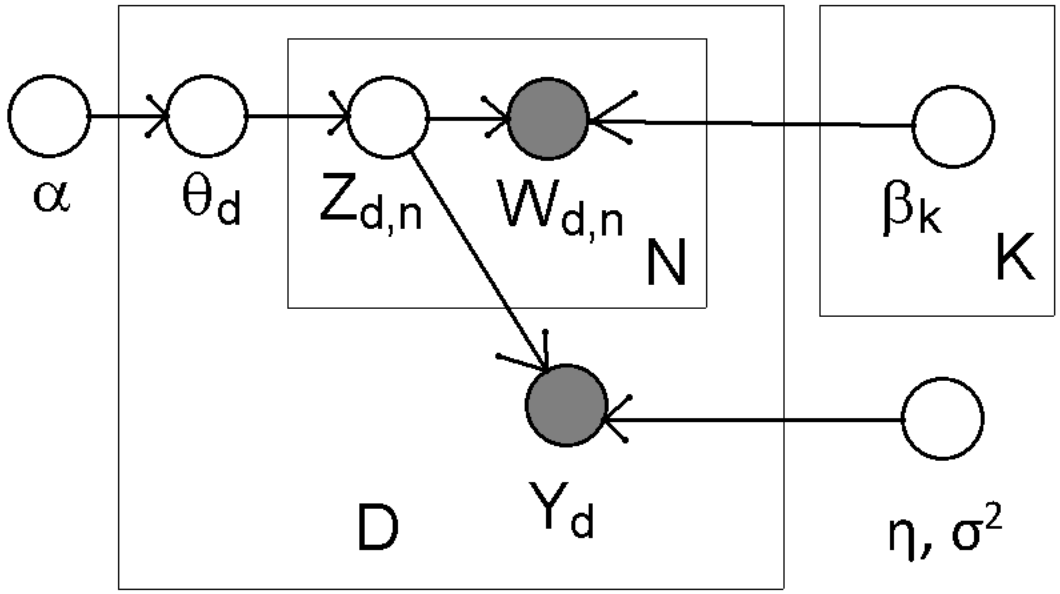

Figure 2: Plate diagram for sLDA.

SLDA IMPLEMENTATION

We applied the sLDA methodology using the **R** package *lda*, with sample code available in *demo(slda)*. First, *demo(slda)* uses the main iterative function *slda.em* to produce the topic model and the topic assignment results, where "*em*" means variational expectation-maximization — an approximation to the maximum likelihood estimation (Blei and McAuliffe 2007). Next, *demo(slda)* generates the plot of credible intervals for each topic, using the **R** package *ggplot2*. Finally, *demo(slda)* uses *slda.predict* to predict the response variable (category) using the sLDA model and plots the output probability distribution.

In the employee satisfaction dataset, each "document" is an individual's survey response, so the number of documents is $M = 530$, and the $K = 10$ topics refer to scores 1-10. The parameters are set to the default values $\alpha = 1, \eta = 0.1, \sigma^2 = 0.25$. We set 10 iterations for the expectation step and 4 iterations for the maximization step.

Using the results from sLDA, we generate the *top.topic.words* for each score, as in Table 1. The **R** function *top.topic.words* selects the five words with the highest posterior probability to appear in each topic. In mathematical terms, the posterior probability is $P(\text{word j} \mid topic\ i,\ data)$; i.e. the probability of getting word j given topic i and the data.

Higher ratings are associated with positive words, e.g. "challenge" and "opportunity," while lower ratings are associated with negative topics, e.g. "lack

of interest."

Table 1: Selected words (tokens) for each topic and rating.

Rating (Topic)	Selected Words
10	challeng opportun learn dai new
9	team work great staff make
8	role feel current perform career
7	job work **enjoi** too project
6	manag work happi last month
5	work get **enjoi** chang need
4	compani project skill resourc engin
3	life balanc compani work peopl
2	lack interest differ **enjoi** requir
1	time hour lot week take

TEXT MINING RESULTS

The "top" topic words in Table 1 are descriptive, though not 100% accurate. For example, the selected words of rating 10 contain the word "challenge" (or the token "challeng"), but "challenge" is not necessarily a positive word. A response could be "It is a challenge dealing with my horrible boss every day." Nevertheless, Figure 3 shows a positive association between the word "challenge" and higher ratings. The correlation between the rating and the number of "challenge"s per comment is 0.920, and when only the ratings 4-10 are included, the correlation is 0.964. Therefore, "challenge" is still considered a positive word.

For another example, the selected words for ratings 2, 5, and 7 all contain the word "enjoy" (or the token "enjoi") because this word is widely used (77 out of 530 ratings). Figure 4 shows how "enjoy" is used in the dataset, and ratings 7-9 have the most "enjoy"s per comment. On the other hand, rating 4 has a higher number of "enjoy"s than rating 5 on average, but this is due to the low number of records in rating 4. The correlation between the rating and the number of "enjoy"s per comment is 0.745, and when only the ratings 4-10 are included, the correlation is 0.530.

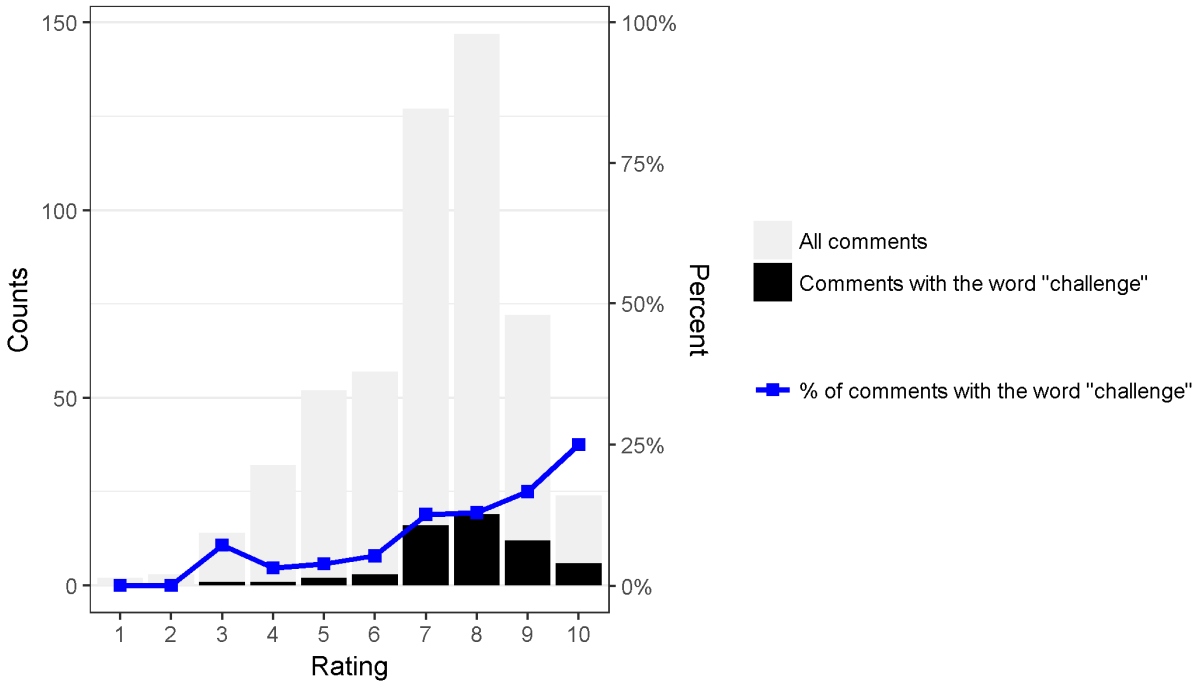

Figure 3: Number of "challenge"s in the comments.

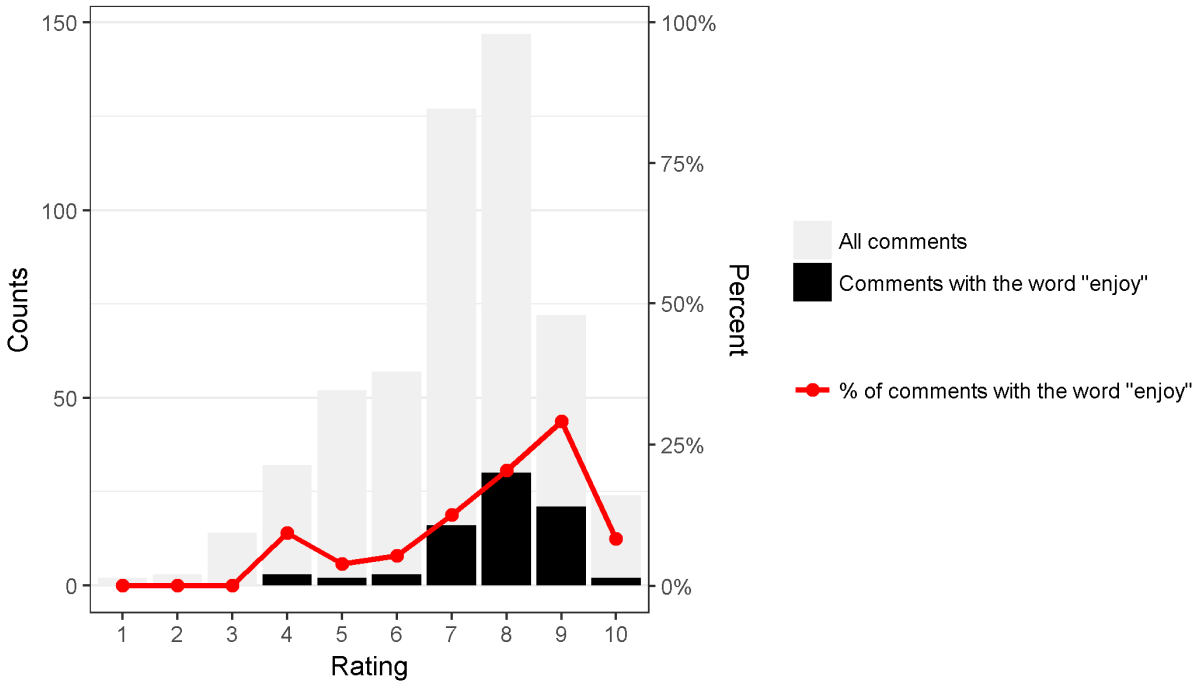

Figure 4: Number of "enjoy"s in the comments.

CREDIBLE INTERVALS OF THE RATINGS FOR EACH TOPIC

The sLDA also outputs credible intervals for the scores associated with each topic (represented by the *top.topic.words*). Figure 5 (Chai 2017) shows the 68% credible intervals (point estimate ± standard error) of the rating estimates. For

instance, given the topic "team work great staff make," the credible interval of the score is approximately between 7 and 8.

The estimated ratings are mostly between 6 and 8, because this is the interval to which the majority of ratings in the data belong. The higher the estimate score, the lighter the interval color is. The t-values (interval thickness) determine statistical significance.

It is also possible to use *slda.predict* to predict the rating of a new document from the sLDA model, but given the small dataset size and the close estimated ratings, we think it is inappropriate to do out-of-sample prediction here.

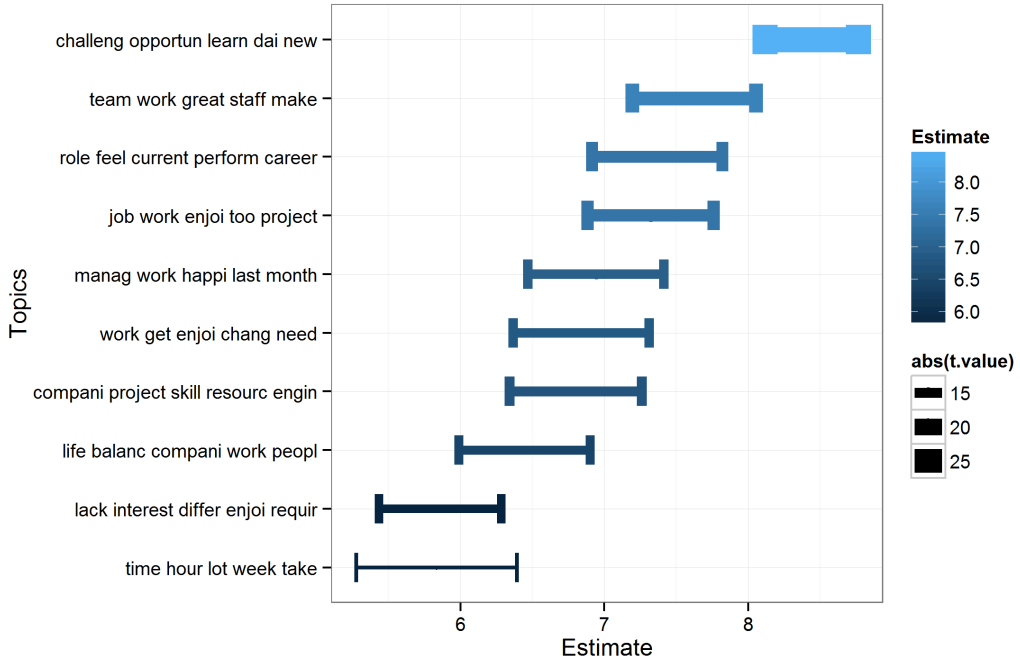

Figure 5: There are 68% credible intervals of the ratings for each topic.

DISCUSSION

Text mining scales well to big data (Martin and Hersh 2014), so automated text mining in surveys is helpful in analyzing large amounts of free text responses. Our survey data on employee satisfaction contain only 530 ratings, but the total number of words in the text already exceeds 10,000. If a large company conducts this type of survey to thousands of their employees, text mining to these responses is essential because it requires more than several hours to manually read through all the text answers.

A common start for automated text mining in surveys is identifying topics for each rating, and we achieved this by applying sLDA on free text responses, making the process much simpler. To implement survey text mining, researchers simply need to run the **R** package *lda* on the survey text, along with the corresponding numerical ratings. A future application is to detect survey

data errors from rating estimations. For example, if a rating is much different than its predicted value from the text comment, the rating may be an error.

Furthermore, survey text mining can also be applied to business intelligence technology, so companies can better understand the customer's needs. One example is to help companies answer questions, such as "Which products are referenced in the survey responses?" and "What topics about the product are people mentioning?" (Chaudhuri, Dayal, and Narasayya 2011). A company that manufactures smart phones may be interested in the results of survey responses on the topics, such as "battery life" and "appearance."

While asking open-ended questions in a survey can potentially increase response rates (O'Cathain and Thomas 2004), researchers should be prepared to analyze the free text responses before collecting the data (Boynton and Greenhalgh 2004). Our straightforward method of survey text-rating analysis allows researchers to be more confident in collecting and analyzing survey text responses, which may lead to publishing better insights from surveys.

This research is a start for potential applications of survey text mining. However, more research is required before moving we would recommend full migration to a supervised automation approach. For instance, the credible intervals are relatively wide due to concentrated ratings in the data. More research into advanced statistical methods and improvements in the sLDA models would be required to narrow the intervals. In addition, the fixed costs of setting up text mining algorithms can be large. Comparative research into the value-add of text and numeric data, as opposed to numeric scale data, is necessary to determine the return on investment. Only after researchers can quantify the benefits and costs will they be able to make an informed decision. Finally, it may be possible for researchers to limit data collection to text data and eliminate numeric scales. The models used in this paper rely on both types of data. Investigation would be necessary to determine whether a text-only approach could be equally or more valuable.

ACKNOWLEDGMENTS

The author would like to thank her PhD advisor at Duke University, David Banks, for his support on this research project. The author would also like to thank Miranda Chung, Brett Moran, Andrew Raim, and Katherine J. Thompson[3] (all from U.S. Census Bureau), for their feedback on the manuscript. The author is also immensely grateful for the comments from the Editor-in-Chief, Ashley Amaya (RTI), and the anonymous reviewers. Last but not least, the author appreciates the copyediting changes from the Copy Editor, Lisa Clancy (CompuScript).

3 The people are listed in alphabetic order of last names.

CONTACT INFORMATION

Author: Christine P. Chai

Email: cpchai21@gmail.com

LinkedIn: https://www.linkedin.com/in/christinechai/

Homepage: https://sites.google.com/site/christinepeijinnchai/

REFERENCES

Bengston, D.N., S.T. Asah, and B.J. Butler. 2011. "The Diverse Values and Motivations of Family Forest Owners in the United States: An Analysis of an Open-Ended Question in the National Woodland Owner Survey." *Small-Scale Forestry* 10 (3): 339–55.

Blei, D.M., and J.D. McAuliffe. 2007. "Supervised Topic Models." In *Advances in Neural Information Processing Systems*, 121–28. Vancouver, Canada: NeurIPS. The paper was presented at NeurIPS 2007 and published in 2008.

Blei, D.M., A.Y. Ng, and M.I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3 (1): 993–1022.

Boynton, P.M., and T. Greenhalgh. 2004. "Selecting, Designing, and Developing Your Questionnaire." *The British Medical Journal* 328 (7451): 1312–15.

Chai, C.P. 2017. "Statistical Issues in Quantifying Text Mining Performance." Doctoral dissertation, Durham, NC: Duke University.

Chang, J. 2015. "Lda: Collapsed Gibbs Sampling Methods for Topic Models. R Package Version 1.4.2." 2015. https://CRAN.R-project.org/package=lda.

Chaudhuri, S., U. Dayal, and V. Narasayya. 2011. "An Overview of Business Intelligence Technology." *Communications of the ACM* 54 (8): 88–98.

Fisher, N.I. 2013. *Analytics for Leaders: A Performance Measurement System for Business Success*. Cambridge, United Kingdom: Cambridge University Press.

Francis, L., and M. Flynn. 2010. "Text Mining Handbook." In *Casualty Actuarial Society E-Forum*, 8. http://www.casact.org/pubs/forum/10spforum/completes10.pdf.

Grimmer, J., and B.M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21 (3): 267–97.

Hughes, M.C. 2010. "Supervised Topic Models for Video Activity Recognition." http://cs.brown.edu/people/mhughes/compsProposal.pdf.

Jackson, K.M., and W.M. Trochim. 2002. "Concept Mapping as an Alternative Approach for the Analysis of Open-Ended Survey Responses." *Organizational Research Methods* 5 (4): 307–36.

Liu, B. 2015. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge, United Kingdom: Cambridge University Press.

Martin, M., and G. Hersh. 2014. *Q&A with a Text Mining pro: How Difficult Is It to Text Mine?* Elsevier Connect. https://www.elsevier.com/connect/q-and-a-with-a-text-mining-pro-how-difficult-is-it-to-text-mine.

O'Cathain, Alicia, and Kate J Thomas. 2004. "'Any Other Comments?' Open Questions on Questionnaires – a Bane or a Bonus to Research?" *BMC Medical Research Methodology* 4 (1). https://doi.org/10.1186/1471-2288-4-25.

Patel, M.D., K.M. Rose, C.R. Owens, H. Bang, and J.S. Kaufman. 2012. "Performance of Automated and Manual Coding Systems for Occupational Data: A Case Study of Historical Records." *American Journal of Industrial Medicine* 55 (3): 228–31. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3316486/.

Porter, M.F. 2001. "Snowball: A Language for Stemming Algorithms." http://snowball.tartarus.org/texts/introduction.html.

Quan, Changqin, Meng Wang, and Fuji Ren. 2014. "An Unsupervised Text Mining Method for Relation Extraction from Biomedical Literature." Edited by Gajendra P. S. Raghava. *PLoS ONE* 9 (7): e102039. https://doi.org/10.1371/journal.pone.0102039.

Ramage, D., D. Hall, R. Nallapati, and C.D. Manning. 2009. "Labeled LDA: A Supervised Topic Model for Credit Attribution in Multi-Labeled Corpora." In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 248–56. Singapore: Association for Computational Linguistics.

Roberts, M.E., B.M. Stewart, D. Tingley, C. Lucas, J. Leder-Luis, S.K. Gadarian, B. Albertson, and D.G. Rand. 2014. "Structural Topic Models for Open-ended Survey Responses." *American Journal of Political Science* 58 (4): 1064–82.

Schuman, H. 2008. *Method and Meaning in Polls and Surveys*. Cambridge MA: Harvard University Press.

Schuman, H., and S. Presser. 1996. *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context*. Thousand Oaks CA: Sage.

Soriano, J., T. Au, and D. Banks. 2013. "Text Mining in Computational Advertising." *Statistical Analysis and Data Mining: The ASA Data Science Journal* 6 (4): 273–85.

Statistical Services Centre. 2001. "Approaches to the Analysis of Survey Data." The University of Reading, United Kingdom. https://www.ilri.org/biometrics/TrainingResources/Documents/University%20of%20Reading/Guides/Guides%20on%20Analysis/ApprochAnalysis.pdf.

SurveyMonkey. n.d. "Analyzing the Data ------ Survey Data Analysis Made Easy." https://www.surveymonkey.com/mp/how-to-analyze-survey-data/.

Wallach, H.M. 2006. "Topic Modeling: Beyond Bag-of-Words." In *Proceedings of the 23rd International Conference on Machine Learning*, 977–84. Pittsburgh, PA: ACM.

Yamanishi, K., and H. Li. 2002. "Mining Open Answers in Questionnaire Data." *IEEE Intelligent Systems* 17 (5): 58–63.

Zhang, Y., R. Jin, and Z.-H. Zhou. 2010. "Understanding Bag-of-Words Model: A Statistical Framework." *International Journal of Machine Learning and Cybernetics* 1 (1–4): 43–52.