Processing Large Quantities of Qualitative Data in Organizational Surveys: Challenges, Solutions, and Applications

Ryan Derickson
Veterans Health Administration National Center for Organization Development

Katerine Osatuke
Veterans Health Administration National Center for Organization Development

Robert Teclaw
Veterans Health Administration National Center for Organization Development

Dee Ramsel
Veterans Health Administration National Center for Organization Development

Qualitative comments provide rich information useful for organizational action planning. Before releasing comments, it is often necessary to scrub them of identifiable information and profanity to maintain the integrity of the survey and protect respondents from repercussions or false accusations. However, processing large quantities of comments (e.g., >50,000) is prohibitively time consuming. We present the method used to develop an automation to filter identifiable information and profanity from comments which greatly expedited the review process and allowed useful applications of comments in the field. We observed high sensitivity, specificity, and accuracy rates comparable to common medical screens with diagnostic utility.

INTRODUCTION

Qualitative comments allow researchers to capture information that is impossible to obtain from quantitative surveys alone. In an organizational-development context, qualitative comments can reveal why employees selected the quantitative responses they did, which is informative for action planning or designing interventions. However, relatively few organizational surveys solicit qualitative responses, especially those distributed to very large (i.e., > 50,000) samples. This underutilization is likely due to the difficulty associated with timely processing of such large volumes of text combined with previous hesitancy to act on comments (Lee & Beres, 2012 SIOP Symposium). Traditional theming methods such as grounded theory (Glaser & Straus 1967, Strauss & Corbin 1998) become increasingly impractical as the number of comments increases, due to the substantial time investment required to identify, refine, and apply themes to the comments.

Strategies exist for mechanically processing text, such as sentiment analysis (e.g., Turney, 2002) and ontology (e.g., Auinger & Fischer, 2008). For example, such analyses may focus on identifying subjective modifiers and their associated nouns (e.g., "my team is excellent") and then assigning a numeric value to the relative strength of the modifier (e.g., in a simplistic example, "good" may be +1, "excellent" may be +2, "terrible" may be -2). Comments with similar positive or negative magnitudes may then be grouped by subject to provide examples of attributes and detriments (Holcomb & Daum, 2012 SIOP Symposium). Additionally, text analysis software such as SPSS Text Mining or SAS Enterprise Miner can apply categorization, regression tree, neural network reasoning, and concept linkage algorithms to identify broad themes and relationships (Crowsey et al., 2007), although such programs do not remove the need for the researcher to actually read the data (Angelique et al., 2005). At the comment or document level, text summarization approaches may attempt to generate a representation of the subject matter and then score sentences on importance or relevance to that representation. Summaries can then be created from selecting and combining the most important sentences (Aggarwak & Zhai, 2012).

Other methods focus on extracting word counts as a means of sense-making, or capturing key words plus a certain amount of text preceding or following the key word (Gerdes Jr, Stringam, & Brookshire, 2008). These and other descriptive analysis methods may be useful for broad sentiment categorization or prediction; for instance, Chevalier and Mayzlin (2006) found an association between the length of an online book review and book sales, and Liu (2006) focused on positive and negative reviews' impact on box office receipts. However, the current state of text processing software does not yet allow the information from comments to be extracted in comprehensive summaries sufficiently accurate or specific for action planning in complex environments such as healthcare.

In order to apply the information from comments towards action planning, the comments need to be shared with organizational stakeholders and particularly with decision makers. Relying on summaries of comments provided by organizational outsiders (e.g. survey contractors) removes the benefit of a direct feedback channel from employees to leaders and introduces the possibility that contractors may inadvertently distort the input from respondents. Summarization by third parties (or software) lacking both the content knowledge and organizational context knowledge could also result in specific and important information being lost through aggregation, which is especially concerning in fields such as healthcare where technical language is common and nuances are important.

The prospect of sharing the survey comments within the organization creates the opportunity for surveyors to preempt several predictable and undesirable consequences of publicly disseminating respondents' comments. For instance, surveyors have the opportunity (and in some cases the responsibility) to encourage that information from comments be used in a manner consistent with the intended purposes of the survey, such as organizational improvement, and not for punitive actions directed at individuals who made unfavorable comments. In order to further reinforce the constructive nature of organizational surveys, surveyors can discourage respondents from using this direct and anonymous communication channel to management as a vehicle for backstabbing, gossip, or unprofessional remarks including profanity. The needs can be only partially met by carefully worded survey instructions. Screening the submitted comments and removing the inappropriate ones (e.g. those remarks that personally attack clearly recognizable individuals), before sharing the results within the organization, may be necessary in order to ensure the constructive nature of the survey feedback. In addition, related concerns from survey respondents (e.g. fears about the potential inappropriate use of their comments, which may potentially compromise the validity of survey responses) also make it crucial that commenters' anonymity (and that of any other person mentioned) is preserved throughout the process of sharing and applying the information from survey comments.

For surveyors to satisfy these data-validity and anonymity concerns, it may often be necessary to filter or redact comments that contain identifying information or profanity. In addition to preserving anonymity, if employees see that comments are not filtered before their release, it may reinforce the use of the survey for less constructive purposes. The time required to complete such a review necessarily increases with the number of comments to the point that it may be prohibitively time consuming for large samples, and the feedback may become outdated by the time it is released. In addition, periodic surveys or multiple comment boxes per survey further exacerbate these requirements. However, if such a review is not done, the organizational value of the survey may be jeopardized, for example as a result of inflammatory comments being released verbatim or as a consequence of surveyors summarizing the comments into overly simplistic themes resulting in information loss and a break of communication between the employee respondents and organizational decision-makers. Given the rich nature of information provided by comments (and lack of sensitive theming software), a more efficient scrubbing process would reduce the labor required and allow more comments to be released as written. However, we are unaware of any software package designed to remove identifying information or profanity from comments "out-of-the-box". Thus far, removing objectionable or identifiable terms has necessitated an exhaustive manual review of comments before their release, which is costly and inefficient for large samples of comments. To address this, we present the method we used to develop and test a program to expedite the scrubbing of large samples of text comments by removing identifying information and profanity through a Visual Basic for Applications program running in Microsoft Excel 2007, and compared the program's ratings to the ratings of trained human raters.

DATA AND RATIONALE

The Department of Veterans Affairs (VA) All Employee Survey (AES) is an annual census of workplace factors such as job satisfaction, civility, and psychological safety. The AES is administered, processed and results are presented by the Veterans Health Administration (VHA) National Center for Organization Development (NCOD). NCOD also facilitates development of action plans at VA organizations that use employee feedback from the AES to design and implement strategies of improving workplace climate and work processes in specific VA locations, service departments, and workgroups.

2012 AES Text Process- Prior to Automation

In the 2012 AES, a comment box was incorporated with the prompt "Please provide any additional comments you may have about your workplace". Comments were limited to 400 characters, and respondents were instructed not to identify any individuals (including themselves), use profanity, or report legal or ethical violations requiring immediate action since the comment may not be seen promptly enough to prevent undesirable outcomes. Respondents were informed that their comments would be provided verbatim to their facility leadership, but would be separated from their quantitative responses in any reports or publicly available datasets, to ease concerns that they may be indirectly linked to their comment from their answers to the AES demographic questions. Of the approximately 180,000 people who answered the quantitative portion of the AES, approximately 58,000 also provided a comment.

Before the comments were made available to facility leadership, it was necessary to remove those that identified specific individuals or contained profanity, as previously discussed. To accomplish this review, NCOD developed the following process. As comments were received from the survey vendor, they were stripped of demographic identifiers except for a code identifying the commenter's facility and workgroup. The code allowed blinding of the raters to the organizational unit of the commenter and also allowed remerging of comments to facility identifiers so that alerts and comments could be sent to the appropriate facilities. Comments were then grouped into files of 500, and each file was sent to two trained raters (chosen randomly from a pool of 20 potential raters) designated as raters A and B. Each comment was independently reviewed by raters A and B, and coded as "No Concern: Share the comment" or "Concern: Do not share the comment". Rater disagreements were reviewed by a third NCOD staff member who rendered a final decision (i.e., "No Concern" or "Concern"). "Concern" comments were defined as those containing: (a) Individual names or other individually identifiable information, or (b) Profanity or offensive language. "Concern" comments were removed from the dataset (comments were dropped rather than redacted because it would have been prohibitively time consuming for raters to perform manual redaction and also manually compare redacted comments to ensure raters A and B redacted the same thing). In addition, an "Alert" rating was assigned to those comments that needed to be addressed immediately regardless of identifying content (e.g., grievances, patient safety issues, ethical concerns, or other sensitive issues), and those comments were conveyed immediately to the leadership of the appropriate organization. All “No Concern” comments were relayed verbatim to facility leadership through text files.

Raters took an average of approximately 2 hours to rate a file of 500 comments, which together with reviewers’ time, produced an estimate of 475 hours required to process 58,000 comments. This is likely a conservative estimate, and does not include substantial rater training time or time spent notifying directors of comments that were flagged as “Alert”.

AUTOMATION DEVELOPMENT METHODOLOGY

Given the inefficiencies of manual review, we investigated whether creating a program to automatically search for identifying information and profanity could produce results satisfactorily similar to human raters. We chose a calibration-validation testing design due to the availability of the large sample of human-rated comments (i.e., the 2012 AES comments). First, we split the file of 58,000 coded comments into halves; the first half was the calibration sample which we used to design and refine the program, and the second half was the validation sample on which we tested the program.

Identifying Terms

Names

The first requirement for the program was a list of names to use as search terms. We obtained a list of the most popular last names, male names, and female names from the Census Bureau (2000 Census), and after removing redundancies (e.g., James is a popular last name and a popular male name) our list contained approximately 10,000 last names, 3,800 female names, and 450 male names. We then obtained a complete list of all VA employees (approximately 40,000 unique first names and 90,000 unique last names) and combined this list with the list of census names. We included Census names to increase the likelihood of catching references to employees who were hired after or left the organization before the full employee name list was compiled. After removing duplicates, 115,117 unique first and last names remained. This list was then divided and sent to NCOD staff members who flagged names in the list that are also common words (e.g., Day, May, Bill). If all names including common words were used, the program would flag virtually every comment and would not create a meaningful workload reduction for human raters. After we reviewed the list of flags, we removed 1,729 names from the search list that are also common words. When deciding whether to remove a common word from the list, the guiding principle used was consideration of whether it seemed at least twice as likely that the term would be used as a word than as a name- if so, the term was removed. This left 113,378 names which comprised our final name list.

Identifiers/Profanity

The second requirement for the program was a list of identifying terms, titles or phrases, and profanity to use as search terms. We started with subsets of 50-100 calibration comments, and manually looked for identifying terms and titles such as “director”, “coordinator”, “ICU chief”, etc. and added these terms to the search list. After we added a number of terms, we wrote the program code in Visual Basic for Applications (due to Excel’s compatibility with SPSS, and SAS which are used to process the quantitative portion of the AES) which rated comments “0” if they did not contain a search term and “1” if they did (for simplicity, at this early stage of development we chose not to address mechanical rating of legal or ethical concerns that humans flagged as “alert” in the 2012 AES processing, which only comprised 0.2% of 2012 comments). We compared the program’s ratings to the human ratings, and if the program rating was a false negative (FN; the program did not flag but humans did), we examined the comment to determine why it was flagged, and added any additional identifying terms to the search list.

We performed many iterations of this process on increasingly large subsets of 100 – 3,000 calibration comments until we reached saturation (i.e., adding more search terms did not meaningfully reduce the FN rate). Throughout this iterative process we remained relatively unconcerned with the false positive rate (FP; the program flagged but humans did not) since this merely represented the program being "overcautious", and humans could presumably review all comments the program drops to determine if they should indeed be dropped or could be retained. The only consideration with the FP rate was that it should be low enough to provide a meaningful reduction to the workload of human raters (if the FP rate is very high, e.g. 75%, then the workload of human raters would be reduced only 25%, and the program's utility would be questionable). We were much more focused on minimizing the FN rate, since it represents identifying information or profanity "getting past" the program and potentially being released.

RESULTS

Calibration

After we had established a thorough list of names, identifying terms, and profanity, we tested the program against the entire calibration sample. The FN rate from this test is an estimation of the lowest FN rate we can expect, since the search term list was built from these comments. Results from this test showed a sensitivity of .927 (true positive / condition positive), specificity of .810 (true negative / condition negative), and accuracy of .815 (N true / N ratings). These results are informative, but the analysis of the validation sample is a better indicator of performance since no comments from this sample were used in building the search term lists.

Validation

Next, we performed a validation test by running the program on the second-half of comments (N = 29,115), which were not considered when creating the search term lists. Results from this validation test are reported in Table 1. Of the FN, 6 contained names, phone numbers, or initials, 106 contained identifying titles or title descriptions, 56 described a situation in enough detail that someone familiar with that facility could deduce who the subjects were, and 5 contained profanity (although it was usually disguised or abbreviated).

TABLE 1
VALIDATION OF RATING PROGRAM ON 2012 AES COMMENTS

		Condition		Performance				
		Positive	Negative	PPV[1]	NPV[2]	Sensitivity[3]	Specificity[4]	Accuracy[5]
Test	Positive	1184	4325	.215	.993	.870	.844	.845
	Negative	177	23429					

[1] Positive predictive value; (true positive / test positive)

[2] Negative predictive value; (true negative / test negative)

[3] True positive / condition positive

[4] True negative / condition negative

[5] (True positive + true negative) / Number of ratings

Additional Validation

Prior to the 2012 AES, the comment box and rating process were piloted on the 2012 Voice of the VA (VoVA) survey, a periodic survey of organizational climate factors similar to the AES. The VoVA response rate is substantially lower than the AES, and 13,345 comments were collected in 2012. These comments were subjected to the same human review process as described for the AES, and the respondent prompt on the VoVA was the same as the prompt on the AES. We tested the program on the entire list of VoVA comments since they represent a completely different sample from a different instrument yet have the same processing requirements (dropping profane or identifying comments) as the AES. Results from this test are presented in Table 2.

TABLE 2
VALIDATION OF RATING PROGRAM ON 2012 VoVA COMMENTS

		Condition		Performance				
		Positive	Negative	PPV[1]	NPV[2]	Sensitivity[3]	Specificity[4]	Accuracy[5]
Test	Positive	531	1699	.238	.992	.861	.867	.866
	Negative	86	11028					

[1] Positive predictive value; (true positive / test positive)

[2] Negative predictive value; (true negative / test negative)

[3] True positive / condition positive

[4] True negative / condition negative

[5] (True positive + true negative) / Number of ratings

IMPLICATIONS

The performance results from the AES and VoVA validation tests were very similar. This seems to indicate that language usage in text comments is relatively stable across survey administrations given the same prompt. This is encouraging and suggests in addition to one-time surveys, the program is effective on cyclical surveys as well which means time savings would compound across survey administrations.

On the AES validation sample, the population prevalence was 4.67% (i.e., 4.67% of all comments contained objectionable material which would have been released were it not for the rating process). After running the program on that sample, the prevalence was reduced to 0.61% (i.e., 0.61% of released comments would contain objectionable material after processing). Said differently, the program flagged 87% of the true positives (sensitivity), ignored 84% of the true negatives (specificity), and correctly classified 85% of all comments (accuracy; Table 1). Although the PPV is relatively low for both the AES and VoVA validations, this is not overly concerning since the false positives were only approximately 15% and 13 % of the respective total number of ratings. This means human raters need review only comments the program flagged (19%) and determine if they were true or false positives, which would result in an 81% workload reduction on the AES validation sample. In the AES validation sample, 96% of incorrectly classified comments were FPs, so human review of this relatively small subset of comments could substantially increase the accuracy rate, and prevent information from being unnecessarily filtered. To bring some context to the accuracy rates, a meta-analysis of 94 studies reported sensitivity ranges of 30% to 87%, and specificity ranges from 86% to 100% for cervical cancer screening tests (Nanda et al., 2000). Menon et al. (2009) reported similar rates for ovarian cancer screening- sensitivity from 75% to 89%, specificity of 98%, and PPV from 2.8% to 43%. Of course the purposes of our program are very different, but it is encouraging that our sensitivity and specificity ranges were well within the ranges of common medical screens with diagnostic utility, especially given that the consequences of incorrect classification are much greater in medical screening than in employee surveys.

If a FN greater than 0% is acceptable, a program created from the methodology we presented has the potential to save substantial resources that would otherwise be required to manually review all comments. In addition to merely rating comments as keep or drop, it is straightforward to have the program redact objectionable terms rather than dropping the whole comment from the file, which allows the content of the comment to be preserved minus the objectionable content. Asking human raters to perform this type of redaction would be even more time-prohibitive and introduce many more opportunities for rater disagreement which would also be time consuming to resolve.

APPLICATIONS

Based on feedback NCOD received regarding the use of comments across VA, it appears that verbatim comments generated valuable discussion that themed or over-summarized comment reports may not have generated. For instance, one hospital discovered employees were misinterpreting items on the quantitative portion of the AES, which will allow clarification on subsequent AES administrations and increased response validity. In another instance, facility leadership collaborated with Union leadership to host a series of town halls to address concerns raised in the comments about specific aspects of communication from leadership to staff. From those discussions, specific and actionable points of process improvement were identified and implemented. Another facility reported cross-walking themes from comments with aggregate scores on the quantitative portion of the AES with leadership and Union representatives. These discussions produced insight into why certain quantitative scores were below expectations, and specific remediation plans were implemented and communicated to staff. These discussions and process improvements illustrate the value provided by incorporating employee comments into action planning, and reinforce the need for distributing comments to the field. Processing automation, such as we presented in this paper, can help make distribution possible by preserving anonymity and professionalism, which becomes increasingly important as the number of comments increases across years.

LIMITATIONS

The main limitation of this methodology is that false negatives are inevitable. There are countless ways a respondent could identify someone (e.g., “the tall blonde man who works in the pharmacy”), and it is impossible to search for every name (including misspellings still humanly recognizable), and it is unfortunate that there is no way, short of exhaustively reading all comments, to know the true FN rate. However, these concerns are mitigated by the reality that even if human raters are employed, the FN rate is almost certainly greater than 0% due to human error, especially when dealing with more than 50,000 comments. Furthermore, even if human raters were to catch every explicit reference to a name or title, the nuances of jobs and language still make it impossible to capture every identifying comment. For instance, a respondent may discuss how their job would be easier if a certain machine was repaired. If that individual was the only person at their facility who worked with that machine, someone from the facility would be able to identify the author but external raters unfamiliar with the facility could not. Or, a respondent may use a particular figure of speech in their comment (e.g., “this policy is bananas!”) that would be locally identifiable if the author often used that figure of speech in conversation, and comments such as these are impossible for humans or machines to identify. Thus, if it is reasonably certain that even human raters will produce a FN greater than 0%, creating an empirically-based program such as the one we presented becomes an attractive option given the high sensitivity, specificity, and accuracy we observed combined with the potential time savings- our program was developed in approximately 75 hours, versus 475 hours spent manually rating comments in 2012. We expect the time savings to compound across years as the program will only require periodic review and revision rather than a complete annual rebuild. The time savings will be even more extreme (and manual review would be even more costly) if additional comment boxes were to be added in future AES administrations.

A similar limitation is the inability of the program to distinguish between terms used as words and terms used as names (e.g., “day” the word, and “Day” the name). If common words were left in the list of names, it would greatly inflate the false positive rate to the point of making the program useless. Therefore, names that are also common words were removed from the search. The definition of “common word” is necessarily subjective, but we followed the general principle of removing a term if it seemed at least twice as likely to be used as a word than as a name given the population and purpose of our survey. For instance, “Alto” is both a word and a name, but given the context of our comments it is unlikely someone would use “alto” in their comment unless they were referring to the proper name, so “Alto” was not removed. Although this introduces the possibility of FN’s via names that are also common words not being flagged, from our calibration and validation tests we found it was much more likely for respondents to identify someone by their title or an identifying situation than by name alone.

GENERALIZABILITY

We see nothing in our iterative process of creating the program and search term lists that would suggest the relatively high sensitivity and specificity rates we observed were unique to our sample. We offer our method and results as a guide to others with similar qualitative comment processing needs, and encourage them to evaluate our methodology against the requirements of their situation and the prevalence of comments with objectionable content in their sample.

ACKNOWLEDGEMENTS

This material is based upon work supported by the Department of Veterans Affairs. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the research investigator(s) and do not necessarily represent the view of the Department of Veterans Affairs or U.S. Government.

REFERENCES

Aggarwal, C. & Zhai, CX (2012). *Mining text data.* Kluwer Academic Publishers: Boston, MA.

Angelique, D., Haughton, D., Nasr, N., Shah, G., Skaletsky, M., & Spack, R. (2005). A review of two text-mining packages: SAS TextMining and WordStat. *The American Statistician, 59,* 89-103.

Auinger, A., & Fischer, M. (2008). Mining consumers’ opinions on the web. *FH Science Day,* 410-419.

Chevalier, J., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research, 3,* 345-354.

Crowsey, M. J., Ramstad, A. R., Gutierrez, D. H., Paladino, G. W., & White, K. P. (2007). An evaluation of unstructured text mining software. Symposium at IEEE Systems and Information Engineering Design Symposium, Charlottesville, VA.

Gerdes Jr., J., Stringam, B., Brookshire, R. (2008). An integrative approach to assess qualitative and quantitative consumer feedback. *Electronic Commerce Research, 8,* 217-234.

Glaser B. & Strauss A. (1967) *The Discovery of Grounded Theory: Strategies for Qualitative Inquiry.* Aldine Publishing Company, Chicago, IL.

Holcomb, K., & Dum, D. (2012). Unleashing the value of analytics to understand employee values. Symposium at Society for Industrial Organizational Psychology (SIOP), San Francisco, CA.

Lee, W., & Beres, R. (2012). New insights from comments using text analytics. Symposium at Society for Industrial Organizational Psychology (SIOP), San Francisco, CA.

Liu, Y. (2006). Word of mouth for movies: its dynamics and impact on box office revenue. *Journal of Marketing, 3,* 74-89.

Menon, U., Gentry-Maharaj, A., Hallett, R., Ryan, A., et al. (2009). Sensitivity and specificity of multimodal and ultrasound screening for ovarian cancer, and stage distribution of detected cancers: Results of the prevalence screen of the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS). *The Lancet Oncology, 10,* 327-340.

Nanda, K., McCrory, D., Myers, E., Bastian, L., Hasselblad, V., Hickey, J., & Matchar, D. (2000). Accuracy of the Papanicolaou test in screening for and follow-up of cervical cytologic abnormalities: A systematic review. *Annals of Internal Medicine, 10,* 810-819.

Strauss, A. L., and Corbin, J. M. (1998), *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Sage Publications, London, UK.

Turney, P. (2002). Thumbs up or thumbs down? Sentiment orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), 417-424.