Entman, R.M. (1993). Framing: Toward Clarification of a Fractured Paradigm. *The Journal of Communication*, 43(4), 51-58.
Hillard, D., Purpura, S. & Wilkerson, J. (2008). Computer-Assisted Topic Classification for Mixed-Methods Social Science Research. *Journal of Information Technology & Politics*, 4(4), 31-46.
Kleinnijenhuis, J. (2008). Reasoning in economic discourse: a network approach to the Dutch Press. In K. Krippendorff & M.A. Bock (Eds.), *The Content Analysis Reader* (pp. 430-442). Thousand Oaks: Sage.
Krippendorff, K. (2004). *Content Analysis*. Thousand Oaks: Sage.
Popping, R. (2000). *Computer-assisted text analysis*. London: Sage.
Reed, D. (2010). Business profile Idea Works: making sense of online chatter. *Columbia Business Times, July 23, http://www.columbiabusinesstimes.com/8490/2010/07/23/*.
Roberts, C. W. (1997). *Text analysis for the social sciences: methods for drawing statistical inferences from texts and transcripts*. Mahwah: Erlbaum.
Stone, P.J., Dunphy, D.C., Smith, M.S. & Ogilvie, D.M. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. Boston: The MIT Press.

Using Computational Techniques to Fill the Gap between Qualitative Data Analysis and Text Analytics

Response to commentary

Curtis Atkisson, Colin Monaghan and Edward Brent

The authors would like to thank both Martine van Selm and Jan Kleinnijenhuis for their thoughtful comments regarding our paper. They each raise a number of interesting points. The comments focus broadly on three areas, and we will structure our response around those three areas. We will first address comments on the literature review, then the Veyor system and finally, comments on the analysis itself.

The literature review

As Van Selm mentions, detailing strengths and weaknesses of different methodologies is a useful tool for understanding when a certain methodology is well suited to address a specific research question or class of research questions. Our review of strengths and weaknesses is not intended to be a comprehensive listing of possible research questions and their relative strengths, but rather to be a basic guide to understanding the differences between three techniques. The point Van Selm makes regarding evaluating the validity of a research method in context of the research question is a good point, but that does not address the statistical validity of the Veyor system, which we evaluate in the paper.

The review of the content analysis literature may have overlooked recent developments in computational techniques used in content analysis (e.g., Krippendorff, 2004), but it does not fail to capture the general lack of emphasis on advanced computational techniques within the field (e.g., Hseih & Shannon, 2005; Mayring, 2000; Stemler, 2001). Regardless of the incorporation of advanced computational techniques into content analysis, there is still methodological space between the three categories. Simply incorporating advanced computational techniques does not alter a reliance on internal validity, reliability and interpretable code clusters which can discourage the discovery of emergent themes and interesting topics mentioned at a low frequency. Specifically, this system allows meaning to emerge, a goal from qualitative data analysis (QDA) and satisfies the desire to examine results statistically, a goal from content analysis.

We are pleased that, aside from the two points addressed above, the authors of the commentaries found the literature review to be well conceived and executed, and that the literature review added to our understanding of the methodological space left between the three described methodologies.

The Veyor system

The comments regarding the Veyor system itself are wide-ranging, and to some extent off-topic. In particular, our goal was to describe a general analysis system that is well designed to fill a place in methodology between QDA, content analysis and text mining. The comments seemed to stray from this goal and focus on issues outside of the scope of the initial paper.

We value the comment made by Van Selm that partitioning the work between a computer and a human is not a new concept. Indeed, for years computers have been programmatically drawing what is presented on your screen, not providing directions to allow you to draw what is on the screen. While this might seem like a basic division of labor, the point is brought up to emphasize that we do not purport to be the first to do this. Instead, we have taken great care to identify the areas in which humans and computers respectively excel. We have tried to use that knowledge to allow a partitioning of work that maximizes the respective efforts. We have found that our unique combination of advanced computational techniques and reduction of cognitive load have increased the effectiveness of this technique.

Kleinnijenhuis references a specific application of Veyor, Globalpoint, and mentions how, in 1989 (Fan & Tims, 1989), sentiment in newspaper articles predicted the outcome of elections. The newspaper article (Reed, 2010) from which Kleinnijenhuis received his information does not indicate that this is a novel idea. Furthermore, the authors made no reference to this article or the Globalpoint system in our initial paper as it was not within the scope of the paper.

The largest part of Kleinnijenhuis's criticism of the description of the Veyor system regards the '"engine" of Veyor. The questions he mentions are very important, and have

indeed occupied thousands of hours of thought while the system was being constructed. Unfortunately, due to the for-profit nature of the system, we are not able to divulge the inner workings. While an ideal world would allow us all to fully describe, and even publish the code of, our system in the public sphere (ala Kleinnijenhuis's Coding Analysis Toolkit), the capital investment of many employees and thousands of hours prevents us from doing that. However, a more technical description of the system might be appropriate here, particularly as to how it responds to comments mentioned in the commentary.

The Veyor system utilizes a suite of computational linguistic strategies in order to allow the construction of expert decision-making systems. These systems allow the program to apply codes automatically. The system utilizes a novel implementation of fuzzy logic to recognize thousands of valid ways of mentioning any particular concept. A semantic web allows the program to identify concepts in context with other concepts, and utilizes the human knowledge that is specified in those semantic webs. Recording units, which was the sentence in this study, are linguistically deconstructed at both the micro (the word) and macro (the grammar of the sentence) levels to allow for applying computational linguistic strategies to the outcome of the linguistic deconstruction. Various statistical techniques (e.g., clustering) are used to both aid in the initial coding process and to assist in refining codes throughout the analysis. This description, while neither exhaustive nor technique specific, should provide an understanding of how the various strategies facilitate an analysis in Veyor, and should answer the questions posed by Kleinnijenhuis in his commentary.

Van Selm brings up a good point about archiving materials from dynamic data sources in the context of research on the internet. This is an important part of conducting research on the web. This, however, was not the issue we sought to solve with the creation of Veyor. Instead, we strove to maximize the analytic abilities of Veyor. As such, the current state of the system requires data to be gathered independently of the system.

A final point regarding the system is the classification of Veyor as a dictionary/thesaurus system by Van Selm. While this may be in line with the definitions laid out by Krippendorf (2004), this does not describe the full functionality of Veyor and it also downplays the complexity of the system. In particular, the Veyor system is not only able to allow for thousands of ways to construct a code, but it also uses expert system type rules to make decisions regarding the application of codes to segments.

The analysis

The majority of the comments regarding the analysis appear to be the result of the constraints of the provided data and the space for the presentation of the analysis. The limitations of the data, whatever they might be, are not a product of data gathering by the authors. Rather the data were provided to us. In addition, space for the final paper was strongly limited, and certain aspects of the analysis needed to be sacrificed. In general,

the process to develop the full coding scheme was not reported in order to allow for a more complete presentation of the results.

A main criticism of Van Selm is in particular need of being addressed. Van Selm asserts that our analysis does not bring much to a qualitative data analysis because it does not elucidate meaning. We strongly disagree with this. Specifically, human inference is typically accomplished by identifying the source of any action, what the cause of the action was and what the consequences were. This combination of information allows us to speculate about meaning regarding the deeper question. This, indeed, is exactly what was accomplished through this analysis. In addition, we sought to bring a naturalistic approach to the question and allow the various actors to indicate causes and consequences, and whether those were positive or negative. By preserving this meaning in the final analysis, we have not only elucidated our understanding of the answer to the question, but we have also allowed actors to provide their own meaning.

Two primary concerns, one each by Kleinnijenhuis and Van Selm, deal with the process of the analysis and appear to be the outcome of limited space for the presentation of the analysis. In particular, Kleinnijenhuis claims that the coding scheme was immutable and a priori while Van Selm criticizes our analysis for being a simple categorization of statements into one of three categories. Both of these are misunderstandings. The coding scheme was constructed using an open coding framework that allowed the data to speak for themselves. All final codes included a minimum of one statement and nearly every statement was coded. Furthermore, the coding was not a simple categorization of whether a statement was regarding an actor, cause or consequence. Instead, the coding was done at a very minute level, allowing for the elucidation of many sub-categories within each main category. Furthermore, many sub-themes were further broken down into constituent parts. We suspect these misinterpretations arose from lack of detail due to the limited space for the presentation of results.

A final category of comments regarded methodological differences in the analysis of social and traditional media. As Kleinnijenhuis mentions, this is an unanswered question. Specifically, how should these different media be weighted in a final analysis? Due to the unresolved nature of the question, and a lack of a theoretical justification for differential weighting, the different media were treated the same in the final analysis. Our final draft of the paper added an analysis regarding differences in those media. The authors are currently working on a study to specifically address that question. Van Selm also mentions a need to accommodate potential methodological differences between the two media. In social media sentiment was easier to observe but context was more difficult. The converse was true with traditional media, with their strong focus on a single topic greatly easing context determination.

Conclusion

A final note in response to the comments is required. This paper sought to present three clearly different ideas: methodological review, presentation of a new software system and analysis of a provided data set. Due to the necessary limitations on space in any journal, this was a daunting task. The methodological review is by no means an exhaustive search of all literature, but it sufficiently integrates different literatures to allow us to see a space between the available methodologies. The presentation of the system was stripped to the bare minimum in order to communicate key functionality while avoiding potential issues regarding intellectual property. Finally, the analysis sought to provide final conclusions from a provided dataset but needed to sacrifice some description of the process in order to accommodate the forum.

References

Fan, D.P. & Tims, A.R. (1989). The impact of the news media on public opinion: American presidential election 1987-1988. *International Journal of Public Opinion Research*, 1, 151-163.

Hseih, H.F. & Shannon, S.E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, 15(1277).

Krippendorff, K. (2004). *Content Analysis*. Thousand Oaks: Sage.

Mayring, P. (2000). Qualitative content analysis. *Forum: Qualitative Social Research*, 1(2).

Reed, D. (2010). Business profile Idea Works: Making sense of online chatter. *Columbia Business Times*, July 23.

Stemler, S. (2001). An overview of content analysis. *Practical Assessment, Research and Evaluation*, 7(17).