The Accuracy in Word of Mouth Valence Classification: Coder Versus Respondents

Cathy Nguyen, Ehrenberg-Bass Institute, UniSA, Cathy.Nguyen@MarketingScience.Info
Jenni Romaniuk, Ehrenberg-Bass Institute, UniSA, Jenni.Romaniuk@MarketingScience.Info

Abstract

Verbatim response coding is becoming increasingly popular within word of mouth (WOM) research, though little is known about the reliability of this approach. We investigate the level of accuracy between a coder's judgment of WOM valence (coded on a 5-point scale) and the valence stated by actual WOM givers. We find a high level of agreement between the valence assigned by the coder and by givers. The inconsistencies were predominantly where comments considered 'slightly positive' by the giver were classified as 'neutral' by the coder. We conclude by suggesting some guidelines on how to improve the measurement and coding of WOM valence.

The Accuracy in Word of Mouth Valence Classification: Coder Versus Respondents

Introduction

Word of mouth (WOM) has been acknowledged as a major influencer in buying behaviour (Arndt, 1967; East, Hammond and Lomax, 2008). The act itself occurs when consumers pass advice or share experiences about products, services or brands (Charlett, Garland and Marr, 1995; East, Hammond and Lomax, 2008). It can entail advising others for or against purchase, or simply be the passing of a positive, negative or neutral statement about an offering (Soderlund and Mattsson, 2009). WOM's *valence,* or 'sentiment', reflects whether it is positive or negative (Buttle, 1998). Generally, it's expected that positive comments or recommendations (PWOM) accelerate brand acceptance and encourage purchase (Traylor and Mathais, 1983), while negative opinions (NWOM) impede brand choice and dissuade purchase (East *et al.*, 2005; Holmes and Lett, 1977). Thus, the valence of the statement expressed can influence the direction of impact of the WOM on receivers. This makes the ability to verify the sentiment of WOM especially important for marketers.

Interest in the effects of WOM has increased dramatically, though the ability to accurately determine its valence, particularly within online contexts, remains a challenge (Jansen *et al.*, 2009; Kim and Hovy, 2004). The Internet forms an attractive landscape for WOM research, as many conversations are publicly available and more easily accessible compared to those that occur offline (Hennig-Thurau *et al.*, 2004). But in spite of such benefits, the online setting does not offer researchers the opportunity to actively prompt participants for the valence of WOM given or received, as could traditionally have been done via surveys. To overcome this, online data providers as well as academics have adopted alternative ways to study sentiment, one of which involves manually coding WOM exactly as the comment appears electronically (Liu, 2006; Godes and Mayzlin, 2004). A problem associated with this approach however, is the difficulty associated with objective, third party, and post hoc classifications of WOM valence (Christiansen and Tax, 2000).

This paper focuses on the measurement issues surrounding manual coding of WOM valence. Past studies in the WOM arena have addressed inter-coder reliability (Godes and Mayzlin, 2004), and more recently, the reliability of automated versus manual coding systems (Jansen *et al.*, 2009). Results from such work typically denote that, despite some limitations, manual and automated coding systems are quite reliable. However, little is understood about the relationship between actual respondent and coder's judgment of WOM valence. This study examines the level of accuracy associated with coders' classification of WOM valence in the context of TV programs. Research that helps establish if and where inconsistencies may occur will provide valuable insight for the abundance of companies and researchers who use manual coding practices in their WOM work. We now review the key literature pertinent to this study.

Literature Review

Interest in WOM valence can be attributed to the influence that positive and NWOM exert on buying behaviour. In the past, the majority of WOM studies where valence is concerned have been conducted using retrospective surveys. As noted by Romaniuk (2007), surveys are ideal for establishing the sentiment of WOM because they allow researchers to actively prompt participants for the valence of the WOM given or received *(e.g., "When was the last time you said something positive or recommend that someone buy X?")*. As it is unnecessary to code for valence, the use of surveys eliminate any potential researcher errors that may arise as a result. In recent years however, WOM researchers as well as marketing practitioners have shifted their attention towards the online arena (Dellarocas, 2003; Godes and Mayzlin, 2004). Not only has the Internet changed the way consumers communicate and exchange WOM, it has also provided marketers with an alternative means for measuring this new phenomenon. The rising number of tools, such as Nielsen Buzz Metrics™, used by businesses to track online WOM content and sentiment about their brands is a testament to the rapid growth in this area.

Two common approaches used to determine WOM valence online are manual coding and automated coding. As researcher-participant interaction is absent, these methods rely on the extraction of raw postings from online sources and the coding of observable content by a third-party. Typically, comments are coded as positive, negative, neutral, mixed, or irrelevant (Godes and Mayzlin, 2004; Liu, 2006), though adaptations to these classifications have also been used (e.g., Jansen et al (2009) used: wretched, bad, so-so, swell and great). Manual coding involves researchers reading and classifying comments into certain valence groups as they see fit. Automated systems however, involves programming technology to automatically detect the presence of positive and negative-type words, and group the cases accordingly into appropriate valence categories (Sebastiani, 2002).

There are a number of limitations associated with both manual and automated coding processes, which lead us to believe that such systems are not as reliable as one would think. As well as being a tedious task (Liu, 2006), accurate content analysis of WOM via manual classification has been shown to be very difficult. For instance, in a study by Godes and Mayzlin (2004), more than 40% of articulations were deemed unclassifiable by coders. The reason for this level of uncertainty is unknown, though the researchers speculate that it may be linked to the subjectivity associated with independent coding. The same study revealed that, of those online comments that could be coded; about 60% achieved agreement in valence classifications between two independent coders. Godes and Mayzlin's (2004) did not discuss the types of comments that received inconsistent classifications.

Coding via electronic devices eliminate the presence of human subjectivity through automatic identification of sentiment. However, a major critique of automated systems is their inability to digest elements of human communication such as humor and sarcasm (Pang and Lee, 2008). Jansen et al. (2009) compared the accuracy of automated versus manual coding of WOM sentiment. They analysed the valence of 150,000 Twitter posts about various firms and found that only 20% of comments contained some expression of sentiment that could be coded. Surprisingly, no significant differences were detected between the manual and automated

approaches. However, the researchers only compared the aggregate distributions of valence groups, ignoring any individual-level inconsistencies that may have occurred.

As outlined, there are numerous options available for classifying WOM valence in research. While past WOM studies have examined the level of consistency between independent coders and between manual and automated systems, little is actually known about how accurate coders' judgments of valence are in relation to the actual giver's intention. This study will shed some light into the appropriateness of manual coding for determining WOM valence. If the use of manual coding in WOM research continues to grow (especially online), then any inaccuracies between coder/respondent need to be addressed. Thus, we present the following questions:

RQ1: How consistent is an independent coder's assessment of valence compared to the respondent's own assessment of valence of WOM given?

RQ2: Are there specific types of WOM that are more prone to ambiguity in valence, and so are difficult for an independent coder to assess?

Research Method

We examine WOM in the context of TV shows- a category known to generate a substantial amount of discussion both off and online (Knowledge Networks, 2009). An online survey was used to collect the relevant data, enabling us to contrast each participant's account of WOM valence and that as assessed by a coder. Respondents were recruited through an Internet panel management company, where individuals opt-in to complete surveys and are rewarded for their time. Our sampling frame included Australian general public respondents aged 18-54 from two Australian capital cities. A quota sampling approach was used to ensure that an equal distribution across gender and age groups was achieved. In total, 616 Australian residents, who had given WOM in the last month, took part in the data collection process. Of those, 35 were removed during the analysis phase due to unusable/invalid responses, leaving a total sample of 581. Hence, over 90% of responses could be coded.

The key measures related to our research include:
Given WOM: To allow for more accurate recollection (Neuman, 2006), participants were asked to draw on the last time they made a comment directed at someone else about a TV show. Over 80% of these had occurred within the last week, with none occurring over one month ago. They were then asked, in an open-ended question, to briefly describe what they said about the show.
Valence as determined by the WOM giver: Respondents indicated how they would describe their comment (from Very Positive to Very Negative). As there is some evidence to suggest that strength of expression influences WOM's impact (East, Hammond and Lomax, 2008), we measure valence using a 5-point scale, as per Jansen et al. (2009).
Valence as determined by the coder: Without exposure to the respondents' ratings, a coder independently assessed all comments and assigned each a valence, using the same 5-point scale.

The valence, as determined by the coder was then compared against the valence as stated by the WOM giver to reveal the level of accuracy between the two parties. In the following section, we present the results.

Results and Discussion

To address RQ1, we compared the distributions of valence groups determined by the respondent and by the coder. We then examined the individual level of accuracy between the two parties. Table 1 shows that the distributions of valence options, determined by the coder, was in line with that stated by the WOM giver. Very Positive was most common for both, and Very Negative was the least common response. There were some discrepancies between respondent-coder in the amount of Very Positive, Slightly Positive and Neutral comments (all $p<0.001$). However, there were no differences in the proportions of Slightly and Very NWOM determined by the coder and respondents ($p>0.05$). We then examined the individual level consistency by calculating the overall proportion of consistent ratings between the respondent and the coder (i.e., of the total 581 comments, what percentage was correctly coded). In total, 62% of coder classifications were consistent with the respondent. This reflects quite a good level of agreement.

Table 1: Overall distributions of WOM valence options

% WOM classification *(n=581)*	**Respondent**	**Coder**
Very Pos.	53	44*
Slightly Pos.	16	10*
Neutral	14	29*
Slightly Neg.	11	11
Very Neg.	6	5

**p<0.05 significant difference between Respondent and Coder (Chi-square)*

To address RQ2, we cross-tabulated the coder's classification of valence with the respondents' classification of valence (Table 2). Overall, the level of accuracy at each valence level was quite high, since over half of comments could be coded correctly. The exception was in the Slightly Positive group, where the majority of responses (79%) were incorrectly coded. This group had the highest level of respondent-coder inconsistency, and this was significantly higher compared to the inconsistencies observed for other valence groups ($p<0.001$). When we explored where the inconsistencies were distributed, we found that around half of comments were classed Neutral by the coder. This could be due to the ambiguous nature of the verbatim responses, where it was often difficult to objectively determine valence (e.g., "*we discussed the last episode", "who got voted off?"*). While the respondent may have discussed the TV show positively, the coder was not given this information and thus, could only assume that the discussion was Neutral. The fact that we rely on respondents' self-iteration of past WOM may be a contributor to this ambiguity.

Table 2: Accuracy across individual WOM valence groups

		Valence by Respondent				
	% Consistency	**Very Pos.** *(n=310)*	**Slightly Pos.** *(n=90)*	**Neutral** *(n=82)*	**Slightly Neg.** *(n=65)*	**Very Neg.** *(n=34)*
Valence by Coder	**Very Pos.**	**73**	24	10	2	3
	Slightly Pos.	10	**21***	7	3	0
	Neutral	16	51	**66**	31	3
	Slightly Neg.	1	3	11	**62**	29
	Very Neg.	0	0	6	3	**65**
	% Consistency	**Positive**		**Neutral**	**Negative**	
	Positive	**74**		17	4	
	Neutral	24		**66**	21	
	Negative	2		17	**75**	

**p<0.05 significant difference between the proportion (%) of consistency versus inconsistency (Chi-square)*

We then examined the level of accuracy across the three primary valence categories by merging Slightly and Very PWOM, and Slightly and Very NWOM (Table 2). In eliminating the 'Slightly' options, the errors associated with these are no longer present. Following the merge, we see that the majority of PWOM was coded positive and NWOM was coded negative. Further, PWOM was rarely coded as Negative and NWOM rarely coded Positive. The error observed for Neutral WOM becomes evenly distributed, and not biased in any direction.

We identified extreme contrary cases where PWOM was coded Negative and vice versa. Given PWOM, which the coder considered Negative, did in fact seem to be genuinely negative (e.g., "*Some drivers are stupid", "TV show has lost its quality").* An explanation for this error could be that respondents made a mistake and chose the wrong valence option. There were four Negative comments, that the coder classified Positive: *"His artwork is really something", "I'm so upset it's over, cancelled!", "Sad it's ending", "The 1st and 2nd episodes were very good but the plot is getting complicated"*. It seems that error was apparent when the coder considered feelings of not wanting a show to end as Positive, yet the WOM givers themselves perceived the feeling of disappointment as Negative. But as this was a rare occurrence, it should not be of major concern.

Conclusions/Implications/Limitations & Future Research

Our research extends on research in the field of WOM valence measurement (Jansen *et al.*, 2009; Pang and Lee, 2008). In this study, over 90% of WOM responses could be coded, as opposed to the lower 20-40% in prior studies (Godes and Mayzlin, 2004; Jansen *et al.*, 2009). Our use of a recall survey rather than online observation is likely the contributor of this difference. The consistency between a coder's evaluations of valence compared with respondents' was quite high at both aggregate and individual level. In most cases, the coder could accurately code PWOM as Positive, and NWOM as Negative. However, there were some inaccuracies, mainly around Slightly PWOM. This suggests that more weakly expressed comments are perhaps harder to evaluate compared to those that are more strongly expressed. The use of three distinct valence groups (Positive, Negative & Neutral) rather than five might benefit from further testing, as it may reduce some errors associated with subjective valence coding. Although, we suggest the use of a 5-point scale where the objective is to investigate WOM *effects* at varying strengths. In order to make the delineation between Slightly P/NWOM and Neutral WOM clearer, we propose substituting the term *Slightly* with *Generally*. This adaptation allows for broader classification and flexibility in that positive comments often also contain some negative element and vice versa. Where recall surveys are used, encouraging respondents to give additional detail in terms of what they said might further aid in the reduction of perceived vagueness of the WOM.

The key limitations are that this is a single study (in one industry so limited scope), with small sample size (especially for NWOM), and the use of one coder. Further, our reliance on respondent recall may be problematic, as we can't be sure that respondents are not rephrasing their original comments due to subjective memory or social desirability bias (Mangold et al., 1999). An extension of this research would be to examine the relationship between intended valence and automated systems. As WOM impact is dependent on the receiver (East, Hammond and Lomax, 2008), it would be useful to compare the coding of valence with those who received WOM rather than givers. Further replication within other categories, with larger sample sizes and the use of multiple coders are highly recommended for future studies.

References

Arndt, J., 1967. Role of product-related conversations in the diffusion of a new product. Journal of Marketing Research 4 (3), 291-295.

Buttle, F., 1998. Word of mouth: understanding and managing referral marketing. Journal of Strategic Marketing 6, 241-254.

Charlett, D., Garland, R. and Marr, N., 1995. How Damaging is Negative Word of Mouth? Marketing Bulletin 6, 42-50.

Christiansen, T. and Tax, S., 2000. Measuring word of mouth: the questions of who and when? Journal of Marketing Communications 6 (3), 185-199.

Dellarocas, C., 2003. The digitization of word-of-mouth: promise and challenges of online reputation mechanisms. Management Science 49 (10), 1407-1424.

East, R., Hammond, K. and Lomax, W., 2008. Measuring the impact of positive and negative word of mouth on brand purchase probability. International Journal of Research in Marketing 25 (3), 215-224.

East, R., Hammond, K., Lomax, W. and Robinson, H., 2005. What is the effect of a recommendation? The Marketing Review 5, 145-157.

Godes, D. and Mayzlin, D., 2004. Using online conversations to study word-of-mouth communication. MARKETING SCIENCE 23 (4), 545-560.

Hennig-Thurau, T., Gwinner, K.P., Walsh, G. and Gremler, D.D., 2004. Electronic word of mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the internet? Journal of Interactive Marketing 18 (1), 38-52.

Holmes, J.H. and Lett, J.D.J., 1977. Product sampling and word of mouth Journal of Advertising Research 17 (5), 35-40.

Jansen, B.J., Zhang, M., Sobel, K. and Chowdury, A., 2009. Twitter power: Tweets as electronic word of mouth. Journal of the American Society for Information Science and Technology 60 (11), 2169-2188.

Kim, S.M. and Hovy, E., 2004. Determining the sentiment of opinions. In: International Conference On Computational Linguistics vol. 4. Geneva, Switzerland

Knowledge Networks (2009) In Press ReleaseKnowledge Networks, pp. 1.

Liu, Y., 2006. Word of mouth for movies: its dynamics and impact on box office revenue. Journal of Marketing 70 (July), 74-89.

Neuman, W.L., 2006. Social research methods: qualitative and quantitative approaches, Allyn & Bacon, Boston.

Pang, B. and Lee, L., 2008. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2 (1-2), 1-135.

Romaniuk, J., 2007. Word of mouth and the viewing of television programs. Journal of Advertising Research 47 (3).

Sebastiani, F., 2002. Machine learning in automated text categorization. ACM Computing Surveys 34 (1), 1-47.

Soderlund, M. and Mattsson, J., 2009. Measuring word-of-mouth activity with recommendation items in service research: What is captured and what is lost? In: ANZMAC Melbourne

Traylor, M. and Mathais, A., 1983. The impact of TV advertising versus word of mouth on the image of lawyers: A projective experiment. Journal of Advertising 12 (4), 42-49.