Impact of Context on Social Media Posts

Will Frankenstein*, Kenneth Joseph, Kathleen M. Carley

Carnegie Mellon University, Pittsburgh, PA
{wfranken, kjoseph, kathleen.carley}@cs.cmu.edu

Abstract. This study examines the role of context in evaluating responses to social media posts online. Current sentiment analysis tools evaluate the content of posts without considering the broader context that the post comes from. Utilizing data from an in-person study, we examine differences between perceived sentiment evaluation when social media response posts are viewed in isolation and perceived sentiment evaluation when social media responses are viewed in the context of the original post. We find that evaluations of responses viewed in context change over 50% of the time. We validate this finding by utilizing simulated data to show the result is not simply a result of data manipulation or noisy data; furthermore, we explore results of this finding with current sentiment analysis tools, examining this result with subsets of our data with high and low kappa values.

Keywords: Twitter, Social Media, Sentiment Analysis, Affect Control Theory

1 Introduction

Traditional approaches to sentiment analysis have three problems: the approaches were originally developed to analyze larger bodies of text, they ignore the social context of social media, and they are primarily focused on only one dimension of sentiment. As social media text can be extremely short, and due to the expense associated with obtaining labeled data necessary to train machine learning algorithms, most approaches to sentiment analysis today rely on extensive lexicons with the goal of having some text match words that we know map to generally positive or negative sentiment [1]-[3].

Most approaches to sentiment analysis in social media focus exclusively on the content of the message, ignoring the metadata and subsequent social context that the message comes out of [4]-[7]. For example, a user posting she is ill will receive positive, supportive posts on social media. Analyzing the social network associated with the flow of those messages would result in an incorrectly classified positive association with that sickness. While some analyses of social network sentiment incorporate analysis of a user's social media ties, these studies rely on aggregated posts and do not consider individual responses to news, topics, or events [8] [9].

Finally, sentiment is typically analyzed along a single dimension: positive and negative, with a minority of research considering objectivity [4] [10]. However, there

are other dimensions to emotions, informed by cultures, which affect how individuals respond to events. Affect control theory (ACT) formalizes the way that individuals respond to events by classifying evaluation, potency, and action, allowing for cross-cultural comparisons of events [11], [12] [13]. Evaluation is the most similar dimension to most sentiment tools today: it is a spectrum from unpleasant and negative to pleasant and positive. Power reflects the social and external relations individuals have, going from weak and powerless to strong and powerful. Activity, in contrast to power, reflects internal relationships to emotion, going from unexciting and inactive to exciting and active. In this study, we utilize a recent dictionary consisting of over 2,000 terms to populate lexicons to identify messages along potency and activity[14].

The paper seeks to explore three key areas: how affect control theory can inform sentiment analysis, how individuals perceive messages seen without context differently from messages with context, and finally, the implications of context for existing tools. We examine the impact of context along all three dimensions of affect control theory, compare evaluations of messages with and without context, and compare individual ratings with automated scores given by sentiment analysis tools.

2 Data

We utilize a subset of a study where 96 individuals collectively rated 5,780 Twitter posts [15]. In the broader study, individuals were given a brief 5-minute training on the three dimensions of ACT, which can be viewed in the technical report [15]. Individuals then each rated 120 Twitter posts three times, once for each dimension of ACT. The 120 Twitter posts evaluated fall into four categories: A) individual Twitter posts, B) responses to Twitter posts, C) the original post that response posts were made to, and D) the same responses seen in category B) – presented this time with the context of the original post. This paper focuses on the changes in response that individuals had from rating category B) tweets to category D) tweets.

Each set of 120 Twitter posts were evaluated twice. We only considered Twitter conversations where the original post was not a response itself. To ensure a broad diversity of topics, we chose Twitter posts from four broad areas, as outlined in the table below.

Table 1. Topic categories for data used.

	Nuclear	**Arab Spring**	**General**	**Haiyan**
Dates	Sep 2014 – Oct 2014	Oct 2009-Nov 2013	Sep 2013 – Aug 2014	Nov 2013 – Dec 2013
Sample Keywords	Nuclear proliferation, uranium	Tahrir Square, Arab Spring	n/a	Haiyan, Typhoon Yolanda
Number of Posts	720	720	720	720

For "General" posts we randomly selected English-language posts from the "Gardenhose", or 10% of the total Twitter firehose, so we did not utilize keywords to select the topics.

3 Comparing responses with and without context

We first explore the data by displaying the distribution of ratings across message categories. We then perform a deeper dive into the different topics making up the dataset and show that we see the same behavior in changed evaluations across all topics. This allows us to make generalizations about the data as a whole and not limited to a subset of our data.

In the histograms below we plot the overall ratings that individuals recorded. Ratings are on a five point Likert scale from negative to positive for Evaluation, weak to strong for Power, and active to passive for Activity. We see that within Power and Activity, the overall profile of responses is consistent whether the post is the original post, the response, or the response viewed with context. The most variation appears to be within Evaluation, which sees slightly more negative posts in responses.

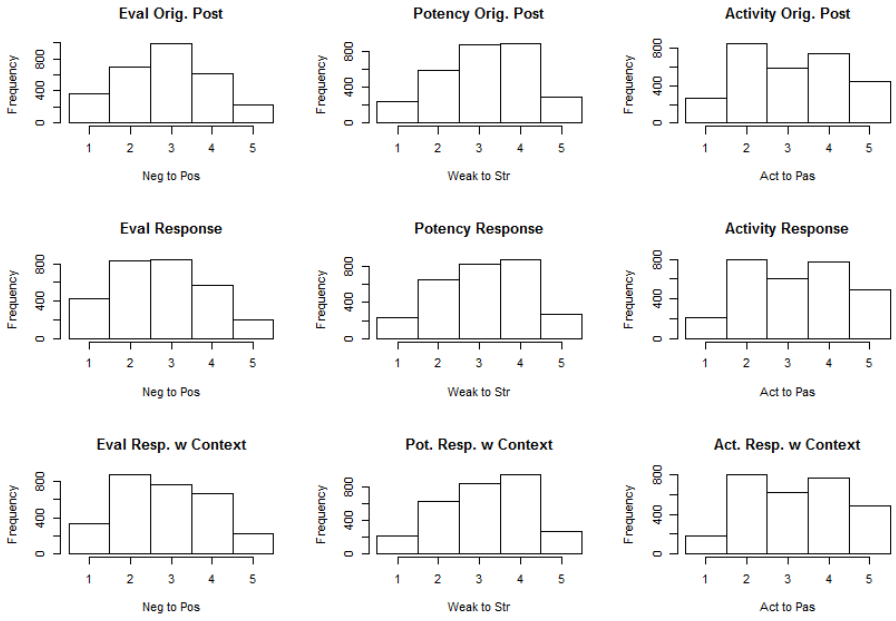

Fig. 1. Histogram of responses across ACT dimensions and post category.

There is some minor variation across topic categories, but there is significant robustness when comparing differences in the evaluation of responses with and without context.

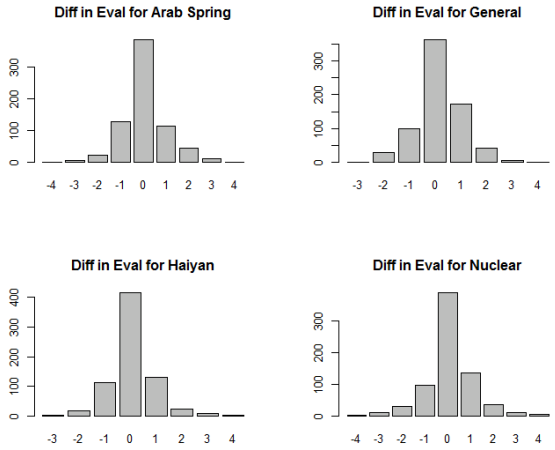

Fig. 2. Difference in evaluation ratings of responses with and without context

We see that in all four categories, we see substantially similar distributions of differences in evaluation across the four categories. The largest bin of changes across all four topics is no change. There is a slightly larger number of individuals changing their evaluations to more negative in Arab Spring tweets.

In repeating this analysis for the other two dimensions of ACT, we see a similar pattern unfold – that regardless of the source of the data, there is a significant amount of change occurring across all three dimensions of Affect Control Theory. We now describe these changes more quantitatively and show that a similar analysis on simulated data does not yield the same result.

4 Features of responses with context

While the histograms give the appearance that the most common change in ratings after seeing context is no change - half the time, individuals are, in fact, changing their ratings. 46% of Evaluations were changed upon seeing context, 50% of Potency ratings were changed, and 52% of Activity ratings were changed.

Table 2. Table of features of changed ratings. Changed Total and Changed Valence percentages are based on all responses; other percentages are based on the number of responses that changed valence.

	Evaluation	Potency	Activity
Changed Total	1,329 (46%)	1,439 (50%)	1497 (52%)
Changed Valence	905 (31%)	1140 (40%)	1138 (40%)
Changed to Neutral	316 (35%)	391 (34%)	360 (32%)
Changed to Pos./Str./Act.	341 (38%)	430 (38%)	375 (33%)
Changed to Neg./Weak/Pas.	267 (30%)	329 (29%)	419 (37%)

In fact, at least 30% of post ratings changed valence after seeing context – 40% for Potency and Activity ratings. Since all ratings were made on a five point Likert scale, we considered all ratings to be one of 3 valences: Negative, Neutral, or Positive for Evaluation; Weak, Neutral or Strong for Potency; and Passive, Neutral, or Active for Activity.

We find that of the posts which changed valence, changes were made relatively uniformly – to either positive/strong/active, neutral, or negative/weak/passive – in overall similar numbers, with about one third of the posts that changed valence going to each category.

We investigated whether viewing context made it more likely to make a post be perceived as being more extreme or whether it largely attenuated ratings. Of posts that changed ratings, 22%, 18%, and 23% of ratings respectively for Evaluation, Potency, and Activity changed to extreme positions. It appears that it is more likely to attenuate an overall rating – while there are larger numbers of neutral ratings in general, a larger proportion of those posts that changed valence across all dimensions of ACT changed to neutral as opposed to changing to a more "extreme" position on the Likert scale.

5 Validation

To validate these findings, we created two simulated datasets with similar summary properties as our data to highlight how the results we obtain are not simply due to data manipulation. Two simulated datasets were used because of uncertainty in the underlying distribution of responses. Each simulated dataset replicates one third of the responses for a given topic area, so there are 12 paired sets of 90 draws.

The first simulated dataset is drawn from a binomial distribution with four draws and a probability of success of 50%. The second simulated dataset is drawn from a multinomial distribution with five bins with probabilities matching the distribution of categories in the Evaluative dataset. As in the original experiment, where we had two

individuals evaluate the same data, we ensured our data had a similar Cohen's kappa of 0.60 by duplicating this data and randomly replacing half of the simulated data.

Table 3. Table of summary statistics comparing binomial and multinomial simulated data

	Eval	Potency	Activity	Binom.	Multi.
1st Quartile	2	2	2	2	2
Median	3	3	3	3	3
Mean	2.8	3.1	3.2	3.0	2.9
3rd Quartile	4	4	4	4	4
Std. Dev.	1.1	1.1	1.2	0.98	1.1

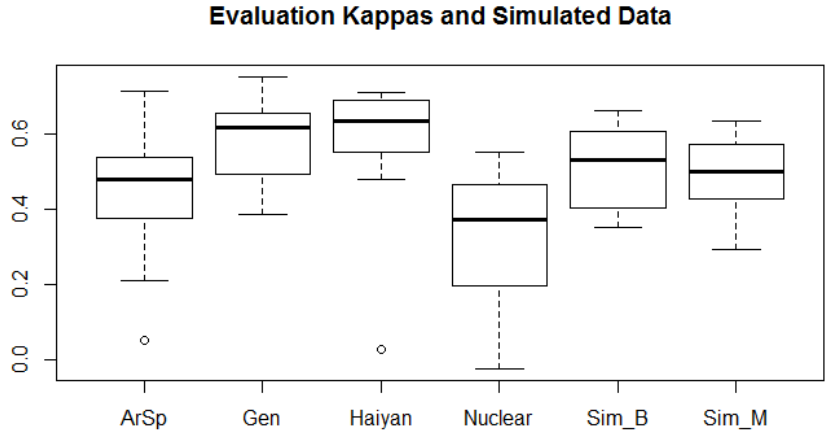

Fig. 3. Distribution of kappas across topic areas and for simulated data; 'Sim_B' indicates data drawn from the binomial distribution, 'Sim_M' indicates data drawn from the multinomial distribution.

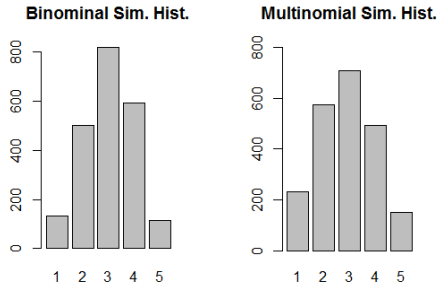

Fig. 4. Histogram of binomial and multinomial simulated data sets.

We find that when comparing our simulated data with difference ratings seen with and without context, the simulated data has a considerably larger variance. In addi-

tion to this larger variance, significantly more respondents choose not to change their rating when compared with our randomly generated data.

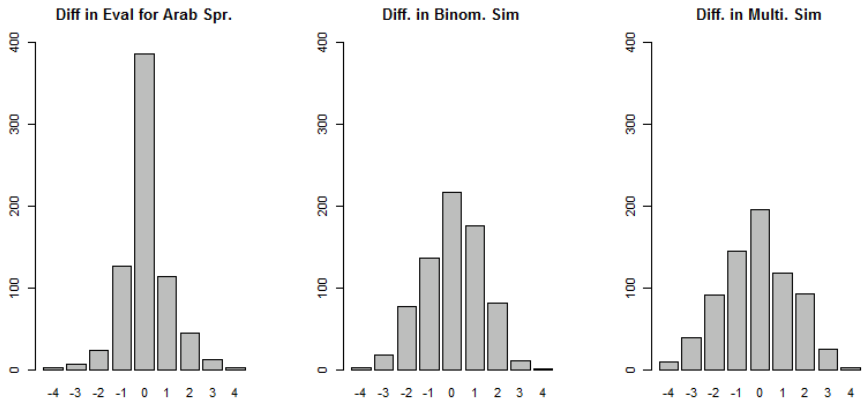

Fig. 5. Histogram of difference in evaluation ratings for Arab Spring contrasted with difference in ratings taken from simulated data.

These results show that a key finding of our study – that about 50% of all ratings change after re-evaluating the message with context – is not simply an artifact of data manipulation.

Table 4. Table of difference statistics, compared with binomial and multinomial simulated data.

	Eval.	**Pot.**	**Act.**	**Binom.**	**Multi.**
Mean	0.10	0.04	0.02	0.03	-0.13
Variance	0.94	1.3	1.5	1.7	2.4
Mean number of 'no change' ratings across topic areas	388	360	346	217	196

6 Implications for current tools

We evaluated the implications of these findings for current sentiment analysis tools in use. We used VADER [16], as well as the most recent ACT lexicon [14] and the CASOS Universal Thesaurus to create a simple sentiment analysis tool that matched n-gram expressions within the Twitter messages – all dictionary methods that are the current standard approach for sentiment analysis tools due to the problem of sparse training data given the short length of Twitter messages [3].

We found through sensitivity analysis that changing the window of what was considered a "neutral" message to being a score from (-0.1,0.1), to (-0.05, 0.05), to (-0.01,

0.01) did not significantly change overall accuracy rates of the sentiment analysis tools used. We set 0.05 as the window for neutral messages for both of the following tables.

Table 5. Sentiment Analysis Tool Matching Rates for Evaluation with neutral score window of 0.05

	VADER	Universal Thesaurus	ACT
Original Message	51%	35%	39%
Response	52%	33%	34%
Response with Context	50%	35%	35%

Table 6. Sentiment Analysis Match rates for Power and Activity using ACT Lexicon, neutral score window of 0.05

	Power	Activity
Original Message	39%	34%
Response	37%	29%
Response with Context	38%	29%

We see that overall sentiment analysis tool ratings appear to match response ratings – as well as original message ratings – at relatively low rates. While our data shows that individuals do change their perceptions of social media messages once they view the message in context, it is harder to draw a connection between automated evaluations of sentiment and what these perceptions are. Future work should further examine the role of size of neutral-rated messages and see if this significantly impacts overall accuracy ratings of sentiment miners.

We take a closer look at match ratings by identifying datasets that had high kappa and datasets that had low kappa. We isolated the ten highest and ten lowest kappa ratings for each axis of ACT; in taking our study, raters had different agreement rates for each axis. All subsets incorporated datasets from each topic group. The table below shows the ranges of the kappas for the data analyzed.

Table 7. Ranges of 10 highest and 10 lowest weighted kappas for each ACT axis.

Evaluation		Potency		Activity	
Low	High	Low	High	Low	High
-0.023-0.37	0.66-0.75	-0.33-0.007	0.33-0.49	-0.13-0.042	0.27-0.34

While we would expect a higher match rate for the subset with higher kappas, we find that overall match rates are identical to the overall population. These rates are not significantly improved by looking at the average rating provided by both raters; additionally, they do not change significantly looking at other dimensions of ACT.

Table 8. Match rates for Evaluation tools, contrasting 10 highest and 10 lowest kappa datasets

	Highest Kappas			Lowest Kappas		
	VADER	UT	ACT	VADER	UT	ACT
Original Message	47%	36%	35%	46%	38%	38%
Response	42%	41%	42%	47%	34%	32%
Response w/ Context	40%	44%	40%	47%	33%	34%

7 Discussion

Social media is a dynamic communication medium – useful for a variety of policy applications, from tracking extremist groups to guiding soft power efforts internationally to raising social awareness. Social media messages are inherently social – they are messages that are meant to be shared and disseminated across platforms. In this study, we have limited our analysis to short conversation snippets on Twitter, and we have only examined the text messages contained in those social media posts. However, many platforms also allow embedding more dynamic media – from GIFs to memes to YouTube videos.

Understanding social contagion and the dynamics of social movements requires understanding the context that these movements come out of. Messages are always viewed in context: for example, a popular online hashtag, #NetflixAndChill, while sounding innocuous, refers to a casual sexual encounter – and quickly served as a shibboleth for 'hip' internet users. Understanding the context surrounding the hashtag requires readers to be aware of considerably more than the current 140 characters Twitter allows in messages. If we are going to quantitatively assess these movements and understand how this change is proliferating across social media, we need to develop better tools that can capture and reflect the ratings of individuals reading and responding to these messages.

The implications of this finding on measuring soft power sentiment: additional structural considerations need to be taken when measuring and observing online discussion of topics. While it is useful to aggregate and distinguish social media posts by their immediate sentiment, additional consideration must be taken to couch posts in the structure of online conversation. If there are several unique posts about a topic, it is going to be more informative to do an analysis of the original posts instead of simply analyzing and aggregating responses to the posts, many of which may be a simple endorsement of the original message. While different social media platforms are able to provide different levels of access to their underlying social network structure, future researchers utilizing social media should try and utilize and incorporate that structure into their sentiment analysis and overall assessment of the platform.

8 References

1. B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," *FNT in Information Retrieval*, vol. 2, no. 1, pp. 1–135, 2008.
2. J. Grimmer and B. M. Stewart, "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts," *Political Analysis*, vol. 21, no. 3, pp. 267–297, Jul. 2013.
3. M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, "Sentiment strength detection in short informal text," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 12, pp. 2544–2558, Dec. 2010.
4. A. Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining," presented at the Proceedings of LREC, 2006.
5. J. W. Pennebaker, C. K. Chung, and M. Ireland, "The development and psychometric properties of LIWC2007," LIWC.net, Austin, TX, USA, LIWC2007, 2007.
6. P. J. Stone, *User's Manual for The General Inquirer*. MIT Press (MA), 1968.
7. A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of twitter data," pp. 30–38, 2011.
8. D. Davidov, O. Tsur, and A. Rappoport, "Enhanced sentiment learning using twitter hashtags and smileys," *Proceedings of the rd international conference on computational linguistics posters*, pp. 241–249, 2010.
9. A. Bermingham, M. Conway, L. McInerney, N. O'Hare, and A. F. Smeaton, *Combining Social Network Analysis and Sentiment Analysis to Explore the Potential for Online Radicalisation*. IEEE, 2009, pp. 231–236.
10. M. Thelwall, K. Buckley, and G. Paltoglou, "Sentiment strength detection for the social web," *Journal of the American Society for Information Science and Technology*, vol. 63, no. 1, pp. 163–173, Jan. 2012.
11. L. S. Lovin, "Affect control theory: An assessment*," *Journal of Mathematical Sociology*, vol. 13, no. 1, pp. 171–192, Jan. 1987.
12. D. R. Heise, "Affect control theory: Concepts and model," *Journal of Mathematical Sociology*, vol. 13, no. 1, pp. 1–33, Jan. 1987.
13. C. E. Osgood, W. H. May, and M. S. Miron, *Cross-cultural Universals of Affective Meaning*. University of Illinois Press, 1975.
14. L. Smith-Lovin and D. T. Robinson, "Interpreting and Responding to Events in Arabic Culture," Office of Naval Research, Grant N00014-09-1-0556, Aug. 2015.
15. W. Frankenstein, K. Joseph, and K. M. Carley, "Social Media ACTion: SOLO Data Description," CMU-ISR-16-103, Feb. 2016.
16. C. J. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text," presented at the … AAAI Conference on Weblogs and Social Media, 2014.